UNITED STATES PATENT AND TRADEMARK OFFICE

CERTIFICATE OF CORRECTION

PATENT NO. : 11,091,794 B2
APPLICATION NO. : 16/995607
DATED : August 17, 2021
INVENTOR(S) : Yuk-Ming Dennis Lo et al.

It is certified that error appears in the above-identified patent and that said Letters Patent is hereby corrected as shown below:

In the Claims

In Column 115, Line 67 and Column 116, Line 36 of Claim 1, please delete "of the first plurality of data structures corresponding to" and insert --of the first plurality of first data structures corresponding to--

Signed and Sealed this
Fifth Day of October, 2021

Drew Hirshfeld
Performing the Functions and Duties of the Under Secretary of Commerce for Intellectual Property and Director of the United States Patent and Trademark Office

UNITED STATES PATENT AND TRADEMARK OFFICE

UNITED STATES DEPARTMENT OF COMMERCE
United States Patent and Trademark Office
Address: COMMISSIONER FOR PATENTS
P.O. Box 1450
Alexandria, Virginia 22313-1450
www.uspto.gov

APPLICATION NO.	FILING DATE	FIRST NAMED INVENTOR	ATTORNEY DOCKET NO.	CONFIRMATION NO.
16/995,607	08/17/2020	Yuk-Ming Dennis Lo	080015-028410US-1199353	1256

20350 7590 09/15/2021

Kilpatrick Townsend & Stockton LLP - West Coast
Mailstop: IP Docketing - 22
1100 Peachtree Street
Suite 2800
Atlanta, GA 30309

EXAMINER
GODDARD, LAURA B

ART UNIT	PAPER NUMBER
1642	

NOTIFICATION DATE	DELIVERY MODE
09/15/2021	ELECTRONIC

Please find below and/or attached an Office communication concerning this application or proceeding.

The time period for reply, if any, is set in the attached communication.

Notice of the Office communication was sent electronically on above-indicated "Notification Date" to the following e-mail address(es):

KTSDocketing2@kilpatrick.foundationip.com
ipefiling@kilpatricktownsend.com

UNITED STATES PATENT AND TRADEMARK OFFICE

Commissioner for Patents
United States Patent and Trademark Office
P.O. Box 1450
Alexandria, VA 22313-1450
www.uspto.gov

Patent No.: 11091794
Issue Date: 08/17/2021
Appl. No.: 16/995,607
Filed: 08/17/2020

PART (A) RESPONSE FOR CERTIFICATES OF CORRECTION

This is a decision on the Certificate of Correction request filed 17 September 2021.

The request for issuance of Certificate of Correction for the above-identified correction(s) under the provisions of 37 CFR 1.322 and/or 1.323 is hereby:

(Check one)
☑ Approved ☐ Approved in Part ☐ Denied

Comments: ____

PART (B) PETITION UNDER 37 CFR 1.324 OR 37 CFR 1.48

☐ This is a decision on the petition filed ____ to correct inventorship under 37 CFR 1.324.

☐ This is a decision on the request under 37 CFR 1.48, petition filed ____. In view of the fact that the patent has already issued, the request under 37 CFR 1.48 has been treated as a petition to correct inventorship under 37 CFR 1.324.

The petition is hereby: ☐ Granted ☐ Dismissed

Comment: ____

The patented filed is being forwarded to Certificate of Corrections Branch for issuance of a certificate naming only the actual inventor or inventors.

/MISOOK YU/
Supervisory Patent Examiner, Art Unit 1642
Technology Center 1600
Phone: (571)272-0839

Certificates of Correction Branch email: CustomerServiceCoC@uspto.gov CoC Central Phone Number: (703) 756-1814

I hereby certify that this correspondence is being filed via EFS-Web with the United States Patent and Trademark Office on September 3, 2021.

KILPATRICK TOWNSEND & STOCKTON LLP

By: /Heather A. Orosco/
Heather A. Orosco

PATENT
Attorney Docket No.: 080015-028410US-1199353
Client Ref. No.: 19/MED/901 US Reg

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

In re Application of:

Yuk-Ming Dennis Lo et al.

Application No.: 16/995,607

Filed: August 17, 2020

For: DETERMINATION OF BASE MODIFICATIONS OF NUCLEIC ACIDS

Customer No.: 20350

Confirmation No.: 1256

Examiner: Laura B. Goddard

Technology Center/Art Unit: 1642

REQUEST FOR CERTIFICATE OF CORRECTION UNDER 37 CFR §1.322

Commissioner for Patents
P.O. Box 1450
Alexandria, VA 22313-1450

Commissioner:

Pursuant to 37 CFR § 1.322, Applicant submits a Certificate of Correction correcting errors made by the Applicant during prosecution regarding the Applicant/Assignee name. The desired corrections are set forth on the attached form PTO/SB/44.

Applicant believes that the Certificate of Correction fee for correcting an Applicant error is $160.00 (37 C.F.R. 1.20(a)). However, if any additional fee is required for this submission, the Commissioner is authorized to charge Deposit Account No. 20-1430.

Respectfully submitted,

/David D. Hsu/
David D. Hsu
Registration No. 71,659

KILPATRICK TOWNSEND & STOCKTON LLP

DDH:hao

75142117V.1

PTO/SB/44 (09-07)
Approved for use through 01/31/2020. OMB 0651-0033
U.S. Patent and Trademark Office; U.S. DEPARTMENT OF COMMERCE
Under the Paperwork Reduction Act of 1995, no persons are required to respond to a collection of information unless it displays a valid OMB control number.
(Also Form PTO-1050)

UNITED STATES PATENT AND TRADEMARK OFFICE
CERTIFICATE OF CORRECTION

PATENT NO. : 11,091,794

APPLICATION NO.: 16/995,607

ISSUE DATE : August 17, 2021

INVENTOR(S) : Yuk-Ming Dennis Lo; Rossa Wai Kwun Chiu; Kwan Chee Chan; Peiyong Jiang; Suk Hang Cheng; Wenlei Peng; On Yee Tse

It is certified that an error appears or errors appear in the above-identified patent and that said Letters Patent is hereby corrected as shown below:

In Column 115, line 67 and Column 116, line 36 of Claim 1, please delete "of the first plurality of data structures corresponding to" and insert --of the first plurality of first data structures corresponding to--

MAILING ADDRESS OF SENDER (Please do not use Customer Number below):
KILPATRICK TOWNSEND & STOCKTON LLP
Mailstop: IP Docketing – 22
1100 Peachtree Street, Suite 2800, Atlanta, GA 30309

This collection of information is required by 37 CFR 1.322, 1.323, and 1.324. The information is required to obtain or retain a benefit by the public which is to file (and by the USPTO to process) an application. Confidentiality is governed by 35 U.S.C. 122 and 37 CFR 1.14. This collection is estimated to take 1.0 hour to complete, including gathering, preparing, and submitting the completed application form to the USPTO. Time will vary depending upon the individual case. Any comments on the amount of time you require to complete this form and/or suggestions for reducing this burden, should be sent to the Chief Information Officer, U.S. Patent and Trademark Office, U.S. Department of Commerce, P.O. Box 1450, Alexandria, VA 22313-1450. DO NOT SEND FEES OR COMPLETED FORMS TO THIS ADDRESS. **SEND TO: Attention Certificate of Corrections Branch, Commissioner for Patents, P.O. Box 1450, Alexandria, VA 22313-1450.**

If you need assistance in completing the form, call 1-800-PTO-9199 and select option 2.

Privacy Act Statement

The **Privacy Act of 1974 (P.L. 93-579)** requires that you be given certain information in connection with your submission of the attached form related to a patent application or patent. Accordingly, pursuant to the requirements of the Act, please be advised that: (1) the general authority for the collection of this information is 35 U.S.C. 2(b)(2); (2) furnishing of the information solicited is voluntary; and (3) the principal purpose for which the information is used by the U.S. Patent and Trademark Office is to process and/or examine your submission related to a patent application or patent. If you do not furnish the requested information, the U.S. Patent and Trademark Office may not be able to process and/or examine your submission, which may result in termination of proceedings or abandonment of the application or expiration of the patent.

The information provided by you in this form will be subject to the following routine uses:

1. The information on this form will be treated confidentially to the extent allowed under the Freedom of Information Act (5 U.S.C. 552) and the Privacy Act (5 U.S.C 552a). Records from this system of records may be disclosed to the Department of Justice to determine whether disclosure of these records is required by the Freedom of Information Act.
2. A record from this system of records may be disclosed, as a routine use, in the course of presenting evidence to a court, magistrate, or administrative tribunal, including disclosures to opposing counsel in the course of settlement negotiations.
3. A record in this system of records may be disclosed, as a routine use, to a Member of Congress submitting a request involving an individual, to whom the record pertains, when the individual has requested assistance from the Member with respect to the subject matter of the record.
4. A record in this system of records may be disclosed, as a routine use, to a contractor of the Agency having need for the information in order to perform a contract. Recipients of information shall be required to comply with the requirements of the Privacy Act of 1974, as amended, pursuant to 5 U.S.C. 552a(m).
5. A record related to an International Application filed under the Patent Cooperation Treaty in this system of records may be disclosed, as a routine use, to the International Bureau of the World Intellectual Property Organization, pursuant to the Patent Cooperation Treaty.
6. A record in this system of records may be disclosed, as a routine use, to another federal agency for purposes of National Security review (35 U.S.C. 181) and for review pursuant to the Atomic Energy Act (42 U.S.C. 218(c)).
7. A record from this system of records may be disclosed, as a routine use, to the Administrator, General Services, or his/her designee, during an inspection of records conducted by GSA as part of that agency's responsibility to recommend improvements in records management practices and programs, under authority of 44 U.S.C. 2904 and 2906. Such disclosure shall be made in accordance with the GSA regulations governing inspection of records for this purpose, and any other relevant (*i.e.*, GSA or Commerce) directive. Such disclosure shall not be used to make determinations about individuals.
8. A record from this system of records may be disclosed, as a routine use, to the public after either publication of the application pursuant to 35 U.S.C. 122(b) or issuance of a patent pursuant to 35 U.S.C. 151. Further, a record may be disclosed, subject to the limitations of 37 CFR 1.14, as a routine use, to the public if the record was filed in an application which became abandoned or in which the proceedings were terminated and which application is referenced by either a published application, an application open to public inspection or an issued patent.
9. A record from this system of records may be disclosed, as a routine use, to a Federal, State, or local law enforcement agency, if the USPTO becomes aware of a violation or potential violation of law or regulation.

Electronic Patent Application Fee Transmittal

Application Number:	16995607
Filing Date:	17-Aug-2020
Title of Invention:	DETERMINATION OF BASE MODIFICATIONS OF NUCLEIC ACIDS
First Named Inventor/Applicant Name:	Yuk-Ming Dennis Lo
Filer:	David Dah-Wei Hsu/Heather Orosco
Attorney Docket Number:	080015-028410US-1199353

Filed as Large Entity

Filing Fees for Utility under 35 USC 111(a)

Description	Fee Code	Quantity	Amount	Sub-Total in USD($)
Basic Filing:				
Pages:				
Claims:				
Miscellaneous-Filing:				
Petition:				
Patent-Appeals-and-Interference:				
Post-Allowance-and-Post-Issuance:				
CERTIFICATE OF CORRECTION	1811	1	160	160

Description	Fee Code	Quantity	Amount	Sub-Total in USD($)
Extension-of-Time:				
Miscellaneous:				
			Total in USD ($)	**160**

Electronic Acknowledgement Receipt

EFS ID:	43682727
Application Number:	16995607
International Application Number:	
Confirmation Number:	1256
Title of Invention:	DETERMINATION OF BASE MODIFICATIONS OF NUCLEIC ACIDS
First Named Inventor/Applicant Name:	Yuk-Ming Dennis Lo
Customer Number:	20350
Filer:	David Dah-Wei Hsu/Heather Orosco
Filer Authorized By:	David Dah-Wei Hsu
Attorney Docket Number:	080015-028410US-1199353
Receipt Date:	03-SEP-2021
Filing Date:	17-AUG-2020
Time Stamp:	13:51:20
Application Type:	Utility under 35 USC 111(a)

Payment information:

Submitted with Payment	yes
Payment Type	CARD
Payment was successfully received in RAM	$160
RAM confirmation Number	E202193D52064119
Deposit Account	201430
Authorized User	Heather Orosco

The Director of the USPTO is hereby authorized to charge indicated fees and credit any overpayment as follows:

37 CFR 1.16 (National application filing, search, and examination fees)

37 CFR 1.17 (Patent application and reexamination processing fees)

37 CFR 1.19 (Document supply fees)

37 CFR 1.21 (Miscellaneous fees and charges)

File Listing:

Document Number	Document Description	File Name	File Size(Bytes)/ Message Digest	Multi Part /.zip	Pages (if appl.)
1	Request for Certificate of Correction	1199353_Request_for_Certificate_of_Correction.pdf	77663 32c49edf5675c5caeaee2b74700a26a1aa355232	no	1
Warnings:					
Information:					
2	Request for Certificate of Correction	1199353_Certificate_of_Correction.pdf	633769 c979ba5cc5b5bde0b87300c53f6b1dece248d4ef	no	2
Warnings:					
Information:					
3	Fee Worksheet (SB06)	fee-info.pdf	37971 0da9ce304fa067e134b3cbea557bc5a5495fefc4	no	2
Warnings:					
Information:					
		Total Files Size (in bytes):	749403		

This Acknowledgement Receipt evidences receipt on the noted date by the USPTO of the indicated documents, characterized by the applicant, and including page counts, where applicable. It serves as evidence of receipt similar to a Post Card, as described in MPEP 503.

New Applications Under 35 U.S.C. 111

If a new application is being filed and the application includes the necessary components for a filing date (see 37 CFR 1.53(b)-(d) and MPEP 506), a Filing Receipt (37 CFR 1.54) will be issued in due course and the date shown on this Acknowledgement Receipt will establish the filing date of the application.

National Stage of an International Application under 35 U.S.C. 371

If a timely submission to enter the national stage of an international application is compliant with the conditions of 35 U.S.C. 371 and other applicable requirements a Form PCT/DO/EO/903 indicating acceptance of the application as a national stage submission under 35 U.S.C. 371 will be issued in addition to the Filing Receipt, in due course.

New International Application Filed with the USPTO as a Receiving Office

If a new international application is being filed and the international application includes the necessary components for an international filing date (see PCT Article 11 and MPEP 1810), a Notification of the International Application Number and of the International Filing Date (Form PCT/RO/105) will be issued in due course, subject to prescriptions concerning national security, and the date shown on this Acknowledgement Receipt will establish the international filing date of the application.

DocCode - PATENT.GRANT

SCORE Placeholder Sheet for IFW Content

Application Number: 16995607 Document Date: 08/17/2021

The presence of this form in the IFW record indicates that the following document type was received in electronic format on the date identified above. This content is stored in the SCORE database.

Since this was an electronic submission, there is no physical artifact folder, no artifact folder is recorded in PALM, and no paper documents or physical media exist. The TIFF images in the IFW record were created from the original documents that are stored in SCORE.

- Patent Grant with Color/Grayscale Drawings

At the time of document entry (noted above):

- USPTO employees may access SCORE content via DAV or via the SCORE web page.
- External customers may access SCORE content via PAIR using the Supplemental Content tab.

Form Revision Date: March 1, 2019

UNITED STATES PATENT AND TRADEMARK OFFICE

UNITED STATES DEPARTMENT OF COMMERCE
United States Patent and Trademark Office
Address: COMMISSIONER FOR PATENTS
P.O. Box 1450
Alexandria, Virginia 22313-1450
www.uspto.gov

APPLICATION NO.	ISSUE DATE	PATENT NO.	ATTORNEY DOCKET NO.	CONFIRMATION NO.
16/995,607	08/17/2021	11091794	080015-028410US-1199353	1256

20350 7590 07/28/2021

Kilpatrick Townsend & Stockton LLP - West Coast
Mailstop: IP Docketing - 22
1100 Peachtree Street
Suite 2800
Atlanta, GA 30309

ISSUE NOTIFICATION

The projected patent number and issue date are specified above.

Determination of Patent Term Adjustment under 35 U.S.C. 154 (b)
(application filed on or after May 29, 2000)

The Patent Term Adjustment is 0 day(s). Any patent to issue from the above-identified application will include an indication of the adjustment on the front page.

If a Continued Prosecution Application (CPA) was filed in the above-identified application, the filing date that determines Patent Term Adjustment is the filing date of the most recent CPA.

Applicant will be able to obtain more detailed information by accessing the Patent Application Information Retrieval (PAIR) WEB site (http://pair.uspto.gov).

Any questions regarding the Patent Term Extension or Adjustment determination should be directed to the Office of Patent Legal Administration at (571)-272-7702. Questions relating to issue and publication fee payments should be directed to the Application Assistance Unit (AAU) of the Office of Data Management (ODM) at (571)-272-4200.

APPLICANT(s) (Please see PAIR WEB site http://pair.uspto.gov for additional applicants):

Yuk-Ming Dennis Lo, Homantin, Kowloon, Hong Kong SAR, CHINA;
The Chinese University of Hong Kong, Shatin, HONG KONG;
Rossa Wai Kwun Chiu, Shatin, New Territories, Hong Kong SAR, CHINA;
Kwan Chee Chan, Jordan, Hong Kong SAR, CHINA;
Peiyong Jiang, Shatin, New Territories, Hong Kong SAR, CHINA;
Suk Hang Cheng, Fanling, New Territories, Hong Kong SAR, CHINA;
Wenlei Peng, Shatin, New Territories, Hong Kong SAR, CHINA;
On Yee Tse, Fanling, New Territories, Hong Kong SAR, CHINA;

The United States represents the largest, most dynamic marketplace in the world and is an unparalleled location for business investment, innovation, and commercialization of new technologies. The USA offers tremendous resources and advantages for those who invest and manufacture goods here. Through SelectUSA, our nation works to encourage and facilitate business investment. To learn more about why the USA is the best country in the world to develop technology, manufacture products, and grow your business, visit SelectUSA.gov.

Under the Paperwork Reduction Act of 1995, no persons are required to respond to a collection of information unless it contains a valid OMB control number.

INFORMATION DISCLOSURE STATEMENT BY APPLICANT (Not for submission under 37 CFR 1.99)		
	Application Number	16/995,607
	Filing Date	August 17, 2020
	First Named Inventor	Yuk-Ming Dennis Lo
	Art Unit	Not Yet Known
	Examiner Name	Not Yet Known
	Attorney Docket Number	080015-028410US-1199353

U.S. PATENTS

Examiner Initial*	Cite No	Patent Number	Kind Code [1]	Issue Date	Name of Patentee or Applicant of cited Document	Pages, Columns, Lines, Where Relevant Passages or Relevant Figures Appear
	1.	9,116,118	B2	08-25-2015	Turner et al.	
	2.	9,175,338	B2	11-03-2015	Flusberg et al.	
	3.	10,590,484	B2	03-17-2020	Korlach et al.	
	4.	8,467,976	B2	06-18-2013	Lo et al.	
Change(s) applied to document, /B.P./ 5/5/2021	5.	10,468,121	B2	~~06-18-2013~~	~~Lo et al.~~ Kermani et al.	11/2019
	6.	10,465,232	B1	11-05-2019	Wu et al.	
	7.	10,438,691	B2	10-08-2019	Kim et al.	
	8.	9,175,341	B2	11-03-2015	Flusberg et al.	

U.S. PATENT APPLICATION PUBLICATIONS

Examiner Initial*	Cite No	Publication Number	Kind Code [1]	Publication Date	Name of Patentee or Applicant of cited Document	Pages, Columns, Lines, Where Relevant Passages or Relevant Figures Appear
	9.	2011/0183320	A1	07-28-2011	Flusberg et al.	
	10.	2017/0037464	A1	02-09-2017	Turner et al.	
	11.	2017/0233802	A1	08-17-2017	Flusberg et al.	
	12.	2018/0201993	A1	07-19-2018	Turner et al.	
	13.	2019/0024162	A1	01-24-2019	Korlach et al.	
	14.	2017/0159132	A1	06-08-2017	Okino et al.	
	15.	2014/0004511	A1	01-02-2014	Korlach et al.	

FOREIGN PATENT DOCUMENTS

Examiner Initial*	Cite No	Foreign Document Number [3]	Country Code [2]	Kind Code [4]	Publication Date	Name of Patentee or Applicant of cited Document	Pages, Columns, Lines where Relevant Passages or Relevant Figures Appear	T [5]
	16.	101896167	KR	B1	09-07-2018		English abstract attached. Equivalent to WO 2009013492 A1 below.	
	17.	2009013492	WO	A1	01-29-2009	The Chinese University of Hong Kong	Equivalent to KR 101896167 B1 above.	

NON-PATENT LITERATURE DOCUMENTS

Examiner Initials*	Cite No	Include name of the author (in CAPITAL LETTERS), title of the article (when appropriate), title of the item (book, magazine, journal, serial, symposium, catalog, etc.), date, page(s), volume-issue number(s), publisher, city and/or country where published.	T [5]

ALL REFERENCES CONSIDERED EXCEPT WHERE LINED THROUGH. /LBG/

Doc Code: MFEE.C.AD

Document Description: Maintenance Fee Address Change

PTO/SB/47 (10-18)
Approved for use through 09/30/2021. OMB 0651-0016
U.S. Patent and Trademark Office; U. S. DEPARTMENT OF COMMERCE
Under the Paperwork Reduction Act of 1995, no persons are required to respond to a collection of information unless it displays a valid OMB control number.

"FEE ADDRESS" INDICATION FORM

Address to:
Mail Stop M Correspondence
Commissioner for Patents
P.O. Box 1450
Alexandria, VA 22313-1450

- OR -

Fax to:
571-273-6500

INSTRUCTIONS: The issue fee must have been paid for application(s) listed on this form. In addition, only an address represented by a Customer Number can be established as the fee address for maintenance fee purposes (hereafter, fee address). A fee address should be established when correspondence related to maintenance fees should be mailed to a different address than the correspondence address for the application. **When to check the first box below**: If you have a Customer Number to represent the fee address. **When to check the second box below**: If you have no Customer Number representing the desired fee address, in which case a completed Request for Customer Number (PTO/SB/125) must be attached to this form. For more information on Customer Numbers, see the Manual of Patent Examining Procedure (MPEP) § 403.

For the following listed application(s), please recognize as the "Fee Address" under the provisions of 37 CFR 1.363 the address associated with:

☑ Customer Number: 00197

OR

☐ The attached Request for Customer Number (PTO/SB/125) form.

PATENT NUMBER (if known)	APPLICATION NUMBER
	16/995,607

Completed by (check one):

☐ Applicant/Inventor

☑ Attorney or Agent of record 71,659 (Reg. No.)

☐ Assignee of record of the entire interest. See 37 CFR 3.71. Statement under 37 CFR 3.73(b) (Form PTO/SB/96) is enclosed or was filed on _____ (Date)

/David D. Hsu/
Signature

David D. Hsu
Typed or printed name

303-571-4000
Requester's telephone number

7/6/2021
Date

NOTE: Signatures of all the inventors or assignees of record of the entire interest or their representative(s) are required. If the assignee is a juristic entity, this form must be signed by a patent practitioner (attorney or agent) of record. Submit multiple forms if more than one signature is required, see below*.

☑ * Total of 1 forms are submitted.

This collection of information is required by 37 CFR 1.363. The information is required to obtain or retain a benefit by the public which is to file (and by the USPTO to process) an application. Confidentiality is governed by 35 U.S.C. 122 and 37 CFR 1.11 and 1.14. This collection is estimated to take 5 minutes to complete, including gathering, preparing, and submitting the completed application form to the USPTO. Time will vary depending upon the individual case. Any comments on the amount of time you require to complete this form and/or suggestions for reducing this burden, should be sent to the Chief Information Officer, U.S. Patent and Trademark Office, U.S. Department of Commerce, P.O. Box 1450, Alexandria, VA 22313-1450. DO NOT SEND COMPLETED FORMS TO THIS ADDRESS. **SEND TO: Mail Stop M Correspondence, Commissioner for Patents, P.O. Box 1450, Alexandria, VA 22313-1450.**

If you need assistance in completing the form, call 1-800-PTO-9199 and select option 2.

Privacy Act Statement

The **Privacy Act of 1974 (P.L. 93-579)** requires that you be given certain information in connection with your submission of the attached form related to a patent application or patent. Accordingly, pursuant to the requirements of the Act, please be advised that: (1) the general authority for the collection of this information is 35 U.S.C. 2(b)(2); (2) furnishing of the information solicited is voluntary; and (3) the principal purpose for which the information is used by the U.S. Patent and Trademark Office is to process and/or examine your submission related to a patent application or patent. If you do not furnish the requested information, the U.S. Patent and Trademark Office may not be able to process and/or examine your submission, which may result in termination of proceedings or abandonment of the application or expiration of the patent.

The information provided by you in this form will be subject to the following routine uses:

1. The information on this form will be treated confidentially to the extent allowed under the Freedom of Information Act (5 U.S.C. 552) and the Privacy Act (5 U.S.C 552a). Records from this system of records may be disclosed to the Department of Justice to determine whether disclosure of these records is required by the Freedom of Information Act.
2. A record from this system of records may be disclosed, as a routine use, in the course of presenting evidence to a court, magistrate, or administrative tribunal, including disclosures to opposing counsel in the course of settlement negotiations.
3. A record in this system of records may be disclosed, as a routine use, to a Member of Congress submitting a request involving an individual, to whom the record pertains, when the individual has requested assistance from the Member with respect to the subject matter of the record.
4. A record in this system of records may be disclosed, as a routine use, to a contractor of the Agency having need for the information in order to perform a contract. Recipients of information shall be required to comply with the requirements of the Privacy Act of 1974, as amended, pursuant to 5 U.S.C. 552a(m).
5. A record related to an International Application filed under the Patent Cooperation Treaty in this system of records may be disclosed, as a routine use, to the International Bureau of the World Intellectual Property Organization, pursuant to the Patent Cooperation Treaty.
6. A record in this system of records may be disclosed, as a routine use, to another federal agency for purposes of National Security review (35 U.S.C. 181) and for review pursuant to the Atomic Energy Act (42 U.S.C. 218(c)).
7. A record from this system of records may be disclosed, as a routine use, to the Administrator, General Services, or his/her designee, during an inspection of records conducted by GSA as part of that agency's responsibility to recommend improvements in records management practices and programs, under authority of 44 U.S.C. 2904 and 2906. Such disclosure shall be made in accordance with the GSA regulations governing inspection of records for this purpose, and any other relevant (*i.e.*, GSA or Commerce) directive. Such disclosure shall not be used to make determinations about individuals.
8. A record from this system of records may be disclosed, as a routine use, to the public after either publication of the application pursuant to 35 U.S.C. 122(b) or issuance of a patent pursuant to 35 U.S.C. 151. Further, a record may be disclosed, subject to the limitations of 37 CFR 1.14, as a routine use, to the public if the record was filed in an application which became abandoned or in which the proceedings were terminated and which application is referenced by either a published application, an application open to public inspection or an issued patent.
9. A record from this system of records may be disclosed, as a routine use, to a Federal, State, or local law enforcement agency, if the USPTO becomes aware of a violation or potential violation of law or regulation.

Electronic Patent Application Fee Transmittal

Application Number:	16995607
Filing Date:	17-Aug-2020
Title of Invention:	DETERMINATION OF BASE MODIFICATIONS OF NUCLEIC ACIDS
First Named Inventor/Applicant Name:	Yuk-Ming Dennis Lo
Filer:	David Dah-Wei Hsu/Heather Orosco
Attorney Docket Number:	080015-028410US-1199353

Filed as Large Entity

Filing Fees for Utility under 35 USC 111(a)

Description	Fee Code	Quantity	Amount	Sub-Total in USD($)
Basic Filing:				
Pages:				
Claims:				
Miscellaneous-Filing:				
Petition:				
Patent-Appeals-and-Interference:				
Post-Allowance-and-Post-Issuance:				
UTILITY APPL ISSUE FEE	1501	1	1200	1200

Description	Fee Code	Quantity	Amount	Sub-Total in USD($)
Extension-of-Time:				
Miscellaneous:				
			Total in USD ($)	**1200**

Electronic Acknowledgement Receipt

EFS ID:	43177271
Application Number:	16995607
International Application Number:	
Confirmation Number:	1256
Title of Invention:	DETERMINATION OF BASE MODIFICATIONS OF NUCLEIC ACIDS
First Named Inventor/Applicant Name:	Yuk-Ming Dennis Lo
Customer Number:	20350
Filer:	David Dah-Wei Hsu/Heather Orosco
Filer Authorized By:	David Dah-Wei Hsu
Attorney Docket Number:	080015-028410US-1199353
Receipt Date:	06-JUL-2021
Filing Date:	17-AUG-2020
Time Stamp:	15:54:45
Application Type:	Utility under 35 USC 111(a)

Payment information:

Submitted with Payment	yes
Payment Type	CARD
Payment was successfully received in RAM	$1200
RAM confirmation Number	E202176F55370984
Deposit Account	201430
Authorized User	Heather Orosco

The Director of the USPTO is hereby authorized to charge indicated fees and credit any overpayment as follows:

37 CFR 1.16 (National application filing, search, and examination fees)

37 CFR 1.17 (Patent application and reexamination processing fees)

37 CFR 1.19 (Document supply fees)

37 CFR 1.20 (Post Issuance fees)

37 CFR 1.21 (Miscellaneous fees and charges)

File Listing:

Document Number	Document Description	File Name	File Size(Bytes)/ Message Digest	Multi Part /.zip	Pages (if appl.)
1	Issue Fee Payment (PTO-85B)	1199353_Issue_Fee_Transmittal.pdf	1633911 8d73a98f650fe35280f084ea12465aa75ffa6076	no	1
Warnings:					
Information:					
2	Maintenance Fee Address Change	1199353_Fee_Address_Indication_Form.pdf	148598 d089cdcff9d7889d668b9f35cfbfa9b3e3a878df	no	2
Warnings:					
Information:					
3	Fee Worksheet (SB06)	fee-info.pdf	30458 3a641d9b5fe952e0c97a1045f522dbd8cbbed4a7	no	2
Warnings:					
Information:					
		Total Files Size (in bytes):	1812967		

This Acknowledgement Receipt evidences receipt on the noted date by the USPTO of the indicated documents, characterized by the applicant, and including page counts, where applicable. It serves as evidence of receipt similar to a Post Card, as described in MPEP 503.

New Applications Under 35 U.S.C. 111

If a new application is being filed and the application includes the necessary components for a filing date (see 37 CFR 1.53(b)-(d) and MPEP 506), a Filing Receipt (37 CFR 1.54) will be issued in due course and the date shown on this Acknowledgement Receipt will establish the filing date of the application.

National Stage of an International Application under 35 U.S.C. 371

If a timely submission to enter the national stage of an international application is compliant with the conditions of 35 U.S.C. 371 and other applicable requirements a Form PCT/DO/EO/903 indicating acceptance of the application as a national stage submission under 35 U.S.C. 371 will be issued in addition to the Filing Receipt, in due course.

New International Application Filed with the USPTO as a Receiving Office

If a new international application is being filed and the international application includes the necessary components for an international filing date (see PCT Article 11 and MPEP 1810), a Notification of the International Application Number and of the International Filing Date (Form PCT/RO/105) will be issued in due course, subject to prescriptions concerning national security, and the date shown on this Acknowledgement Receipt will establish the international filing date of the application.

PART B - FEE(S) TRANSMITTAL

Complete and send this form, together with applicable fee(s), by mail or fax, or via EFS-Web.

By mail, send to: Mail Stop ISSUE FEE
Commissioner for Patents
P.O. Box 1450
Alexandria, Virginia 22313-1450

By fax, send to: (571)-273-2885

INSTRUCTIONS: This form should be used for transmitting the ISSUE FEE and PUBLICATION FEE (if required). Blocks 1 through 5 should be completed where appropriate. All further correspondence including the Patent, advance orders and notification of maintenance fees will be mailed to the current correspondence address as indicated unless corrected below or directed otherwise in Block 1, by (a) specifying a new correspondence address; and/or (b) indicating a separate "FEE ADDRESS" for maintenance fee notifications.

CURRENT CORRESPONDENCE ADDRESS (Note: Use Block 1 for any change of address)

20350 7590 04/07/2021
Kilpatrick Townsend & Stockton LLP - West Coast
Mailstop: IP Docketing - 22
1100 Peachtree Street
Suite 2800
Atlanta, GA 30309

Note: A certificate of mailing can only be used for domestic mailings of the Fee(s) Transmittal. This certificate cannot be used for any other accompanying papers. Each additional paper, such as an assignment or formal drawing, must have its own certificate of mailing or transmission.

Certificate of Mailing or Transmission

I hereby certify that this Fee(s) Transmittal is being deposited with the United States Postal Service with sufficient postage for first class mail in an envelope addressed to the Mail Stop ISSUE FEE address above, or being transmitted to the USPTO via EFS-Web or by facsimile to (571) 273-2885, on the date below.

Heather A. Orosco	(Typed or printed name)
/Heather A. Orosco/ via EFS-Web	(Signature)
July 6, 2021	(Date)

APPLICATION NO.	FILING DATE	FIRST NAMED INVENTOR	ATTORNEY DOCKET NO.	CONFIRMATION NO.
16/995,607	08/17/2020	Yuk-Ming Dennis Lo	080015-028410US-1199353	1256

TITLE OF INVENTION: DETERMINATION OF BASE MODIFICATIONS OF NUCLEIC ACIDS

APPLN. TYPE	ENTITY STATUS	ISSUE FEE DUE	PUBLICATION FEE DUE	PREV. PAID ISSUE FEE	TOTAL FEE(S) DUE	DATE DUE
nonprovisional	~~SMALL~~ UNDISCOUNTED	~~$600~~ $1200	$0.00	$0.00	~~$600~~ $1200	07/07/2021

EXAMINER	ART UNIT	CLASS-SUBCLASS
GODDARD, LAURA B	1642	435-006110

1. Change of correspondence address or indication of "Fee Address" (37 CFR 1.363).

☐ Change of correspondence address (or Change of Correspondence Address form PTO/SB/122) attached.

☑ "Fee Address" indication (or "Fee Address" Indication form PTO/SB/47; Rev 03-09 or more recent) attached. **Use of a Customer Number is required.**

2. For printing on the patent front page, list
(1) The names of up to 3 registered patent attorneys or agents OR, alternatively,
(2) The name of a single firm (having as a member a registered attorney or agent) and the names of up to 2 registered patent attorneys or agents. If no name is listed, no name will be printed.

1 Kilpatrick Townsend & Stockton LLP
2
3

3. ASSIGNEE NAME AND RESIDENCE DATA TO BE PRINTED ON THE PATENT (print or type)

PLEASE NOTE: Unless an assignee is identified below, no assignee data will appear on the patent. If an assignee is identified below, the document must have been previously recorded, or filed for recordation, as set forth in 37 CFR 3.11 and 37 CFR 3.81(a). Completion of this form is NOT a substitute for filing an assignment.

(A) NAME OF ASSIGNEE	(B) RESIDENCE: (CITY and STATE OR COUNTRY)
The Chinese University of Hong Kong	Shatin, New Territories, Hong Kong

Please check the appropriate assignee category or categories (will not be printed on the patent) : ☐ Individual ☑ Corporation or other private group entity ☐ Government

4a. Fees submitted: ☑ Issue Fee ☐ Publication Fee (if required) ☐ Advance Order - # of Copies _____

4b. Method of Payment: *(Please first reapply any previously paid fee shown above)*

☑ Electronic Payment via EFS-Web ☐ Enclosed check ☐ Non-electronic payment by credit card (Attach form PTO-2038)

☑ The Director is hereby authorized to charge the required fee(s), any deficiency, or credit any overpayment to Deposit Account No. 20-1430

5. **Change in Entity Status** (from status indicated above)

☐ Applicant certifying micro entity status. See 37 CFR 1.29

☐ Applicant asserting small entity status. See 37 CFR 1.27

☒ Applicant changing to regular undiscounted fee status.

NOTE: Absent a valid certification of Micro Entity Status (see forms PTO/SB/15A and 15B), issue fee payment in the micro entity amount will not be accepted at the risk of application abandonment.
NOTE: If the application was previously under micro entity status, checking this box will be taken to be a notification of loss of entitlement to micro entity status.
NOTE: Checking this box will be taken to be a notification of loss of entitlement to small or micro entity status, as applicable.

NOTE: This form must be signed in accordance with 37 CFR 1.31 and 1.33. See 37 CFR 1.4 for signature requirements and certifications.

Authorized Signature /David D. Hsu/ Date 7/6/2021

Typed or printed name David D. Hsu Registration No. 71,659

UNITED STATES PATENT AND TRADEMARK OFFICE

UNITED STATES DEPARTMENT OF COMMERCE
United States Patent and Trademark Office
Address: COMMISSIONER FOR PATENTS
P.O. Box 1450
Alexandria, Virginia 22313-1450
www.uspto.gov

APPLICATION NO.	FILING DATE	FIRST NAMED INVENTOR	ATTORNEY DOCKET NO.	CONFIRMATION NO.
16/995,607	08/17/2020	Yuk-Ming Dennis Lo	080015-028410US-1199353	1256

20350 7590 05/26/2021

Kilpatrick Townsend & Stockton LLP - West Coast
Mailstop: IP Docketing - 22
1100 Peachtree Street
Suite 2800
Atlanta, GA 30309

EXAMINER
GODDARD, LAURA B

ART UNIT	PAPER NUMBER
1642	

NOTIFICATION DATE	DELIVERY MODE
05/26/2021	ELECTRONIC

Please find below and/or attached an Office communication concerning this application or proceeding.

The time period for reply, if any, is set in the attached communication.

Notice of the Office communication was sent electronically on above-indicated "Notification Date" to the following e-mail address(es):

KTSDocketing2@kilpatrick.foundationip.com
ipefiling@kilpatricktownsend.com

Notice of Allowability	Application No. 16/995,607	Applicant(s) Lo et al.	
	Examiner LAURA B GODDARD	Art Unit 1642	AIA (FITF) Status Yes

-- The MAILING DATE of this communication appears on the cover sheet with the correspondence address--

All claims being allowable, PROSECUTION ON THE MERITS IS (OR REMAINS) CLOSED in this application. If not included herewith (or previously mailed), a Notice of Allowance (PTOL-85) or other appropriate communication will be mailed in due course. **THIS NOTICE OF ALLOWABILITY IS NOT A GRANT OF PATENT RIGHTS.** This application is subject to withdrawal from issue at the initiative of the Office or upon petition by the applicant. See 37 CFR 1.313 and MPEP 1308.

1. ☑ This communication is responsive to 5/20/2021.
   ☐ A declaration(s)/affidavit(s) under **37 CFR 1.130(b)** was/were filed on _____.

2. ☐ An election was made by the applicant in response to a restriction requirement set forth during the interview on _____; the restriction requirement and election have been incorporated into this action.

3. ☑ The allowed claim(s) is/are 1-17 and 70-71 . As a result of the allowed claim(s), you may be eligible to benefit from the **Patent Prosecution Highway** program at a participating intellectual property office for the corresponding application. For more information , please see **http://www.uspto.gov/patents/init_events/pph/index.jsp** or send an inquiry to **PPHfeedback@uspto.gov.**

4. ☐ Acknowledgment is made of a claim for foreign priority under 35 U.S.C. § 119(a)-(d) or (f).

   Certified copies:

   a) ☐ All b) ☐ Some *c) ☐ None of the:

   1. ☐ Certified copies of the priority documents have been received.
   2. ☐ Certified copies of the priority documents have been received in Application No. _____ .
   3. ☐ Copies of the certified copies of the priority documents have been received in this national stage application from the International Bureau (PCT Rule 17.2(a)).

   * Certified copies not received: _____ .

Applicant has THREE MONTHS FROM THE "MAILING DATE" of this communication to file a reply complying with the requirements noted below. Failure to timely comply will result in ABANDONMENT of this application.
THIS THREE-MONTH PERIOD IS NOT EXTENDABLE.

5. ☐ CORRECTED DRAWINGS (as "replacement sheets") must be submitted.
   ☐ including changes required by the attached Examiner's Amendment / Comment or in the Office action of Paper No./Mail Date _____ .

   Identifying indicia such as the application number (see 37 CFR 1.84(c)) should be written on the drawings in the front (not the back) of each sheet. Replacement sheet(s) should be labeled as such in the header according to 37 CFR 1.121(d).

6. ☐ DEPOSIT OF and/or INFORMATION about the deposit of BIOLOGICAL MATERIAL must be submitted. Note the attached Examiner's comment regarding REQUIREMENT FOR THE DEPOSIT OF BIOLOGICAL MATERIAL.

Attachment(s)

1. ☐ Notice of References Cited (PTO-892)
2. ☑ Information Disclosure Statements (PTO/SB/08), Paper No./Mail Date _____.
3. ☐ Examiner's Comment Regarding Requirement for Deposit of Biological Material _____.
4. ☐ Interview Summary (PTO-413), Paper No./Mail Date. _____.
5. ☐ Examiner's Amendment/Comment
6. ☐ Examiner's Statement of Reasons for Allowance
7. ☐ Other _____.

/Laura B Goddard/
Primary Examiner, Art Unit 1642

INFORMATION DISCLOSURE STATEMENT BY APPLICANT (Not for submission under 37 CFR 1.99)		
	Application Number	16/995,607
	Filing Date	August 17, 2020
	First Named Inventor	Yuk-Ming Dennis Lo
	Art Unit	1642
	Examiner Name	Laura B. Goddard
	Attorney Docket Number	080015-028410US-1199353

U.S. PATENTS						
Examiner Initial*	Cite No	Patent Number	Kind Code [1]	Issue Date	Name of Patentee or Applicant of cited Document	Pages, Columns, Lines, Where Relevant Passages or Relevant Figures Appear

U.S. PATENT APPLICATION PUBLICATIONS						
Examiner Initial*	Cite No	Publication Number	Kind Code [1]	Publication Date	Name of Patentee or Applicant of cited Document	Pages, Columns, Lines, Where Relevant Passages or Relevant Figures Appear

FOREIGN PATENT DOCUMENTS								
Examiner Initial*	Cite No	Foreign Document Number[3]	Country Code[2]	Kind Code[4]	Publication Date	Name of Patentee or Applicant of cited Document	Pages, Columns, Lines where Relevant Passages or Relevant Figures Appear	T[5]

NON-PATENT LITERATURE DOCUMENTS			
Examiner Initials*	Cite No	Include name of the author (in CAPITAL LETTERS), title of the article (when appropriate), title of the item (book, magazine, journal, serial, symposium, catalog, etc.), date, page(s), volume-issue number(s), publisher, city and/or country where published.	T[5]
	1.	Examination Report under Section 18(3) dated April 12, 2021 in GB Patent Application No. 2102808.9. 6 pages.	
	2.	Combined Search and Examination Report under Sections 17 & 18(3) dated April 15, 2021 in GB Patent Application No. 2103010.1. 4 pages.	
	3.	HE, Jing et al.; "Tet1 promotes leukemic growth of AML1-ETO+ AML via 5-hydroxymethylcytosine marks and its oncogenic role and high expression can be antagonized using Olaparib, an inhibitor of its binding partner PARP1"; Experimental Hematology; 2017; Volume 53, Supplement; pages S92-S93.	

EXAMINER SIGNATURE			
Examiner Signature	/Laura B Goddard/	Date Considered 05/24/2021	

ALL REFERENCES CONSIDERED EXCEPT WHERE LINED THROUGH. /LBG/

INFORMATION DISCLOSURE STATEMENT BY APPLICANT (Not for submission under 37 CFR 1.99)		
	Application Number	16/995,607
	Filing Date	August 17, 2020
	First Named Inventor	Yuk-Ming Dennis Lo
	Art Unit	1642
	Examiner Name	Laura B. Goddard
	Attorney Docket Number	080015-028410US-1199353

*EXAMINER: Initial if reference considered, whether or not citation is in conformance with MPEP 609. Draw line through a citation if not in conformance and not considered. Include copy of this form with next communication to applicant.

[1] See Kind Codes of USPTO Patent Documents at www.USPTO.GOV or MPEP 901.04. [2] Enter office that issued the document, by the two-letter code (WIPO Standard ST.3). [3] For Japanese patent documents, the indication of the year of the reign of the Emperor must precede the serial number of the patent document.
[4] Kind of document by the appropriate symbols as indicated on the document under WIPO Standard ST.16 if possible. [5] Applicant is to place a check mark here if English language translation is attached.

ALL REFERENCES CONSIDERED EXCEPT WHERE LINED THROUGH. /LBG/

INFORMATION DISCLOSURE STATEMENT BY APPLICANT (Not for submission under 37 CFR 1.99)		
	Application Number	16/995,607
	Filing Date	August 17, 2020
	First Named Inventor	Yuk-Ming Dennis Lo
	Art Unit	1642
	Examiner Name	Laura B. Goddard
	Attorney Docket Number	080015-028410US-1199353

U.S. PATENTS						
Examiner Initial*	Cite No	Patent Number	Kind Code [1]	Issue Date	Name of Patentee or Applicant of cited Document	Pages, Columns, Lines, Where Relevant Passages or Relevant Figures Appear

U.S. PATENT APPLICATION PUBLICATIONS						
Examiner Initial*	Cite No	Publication Number	Kind Code [1]	Publication Date	Name of Patentee or Applicant of cited Document	Pages, Columns, Lines, Where Relevant Passages or Relevant Figures Appear

FOREIGN PATENT DOCUMENTS								
Examiner Initial*	Cite No	Foreign Document Number[3]	Country Code[2]	Kind Code[4]	Publication Date	Name of Patentee or Applicant of cited Document	Pages, Columns, Lines where Relevant Passages or Relevant Figures Appear	T[5]

NON-PATENT LITERATURE DOCUMENTS			
Examiner Initials*	Cite No	Include name of the author (in CAPITAL LETTERS), title of the article (when appropriate), title of the item (book, magazine, journal, serial, symposium, catalog, etc.), date, page(s), volume-issue number(s), publisher, city and/or country where published.	T[5]
	1.	Examination Report under Section 18(3) dated April 12, 2021 in GB Patent Application No. 2102808.9. 6 pages.	
	2.	Combined Search and Examination Report under Sections 17 & 18(3) dated April 15, 2021 in GB Patent Application No. 2103010.1. 4 pages.	
	3.	HE, Jing et al.; "Tet1 promotes leukemic growth of AML1-ETO+ AML via 5-hydroxymethylcytosine marks and its oncogenic role and high expression can be antagonized using Olaparib, an inhibitor of its binding partner PARP1"; Experimental Hematology; 2017; Volume 53, Supplement; pages S92-S93.	

EXAMINER SIGNATURE			
Examiner Signature		Date Considered	

INFORMATION DISCLOSURE STATEMENT BY APPLICANT (Not for submission under 37 CFR 1.99)		
	Application Number	16/995,607
	Filing Date	August 17, 2020
	First Named Inventor	Yuk-Ming Dennis Lo
	Art Unit	1642
	Examiner Name	Laura B. Goddard
	Attorney Docket Number	080015-028410US-1199353

*EXAMINER: Initial if reference considered, whether or not citation is in conformance with MPEP 609. Draw line through a citation if not in conformance and not considered. Include copy of this form with next communication to applicant.

[1] See Kind Codes of USPTO Patent Documents at www.USPTO.GOV or MPEP 901.04. [2] Enter office that issued the document, by the two-letter code (WIPO Standard ST.3). [3] For Japanese patent documents, the indication of the year of the reign of the Emperor must precede the serial number of the patent document.
[4] Kind of document by the appropriate symbols as indicated on the document under WIPO Standard ST.16 if possible. [5] Applicant is to place a check mark here if English language translation is attached.

I hereby certify that this correspondence is being filed via EFS-Web with the United States Patent and Trademark Office on May 20, 2021.

KILPATRICK TOWNSEND & STOCKTON LLP

By: /Heather A. Orosco/
Heather A. Orosco

PATENT
Attorney Docket No.: 080015-028410US-1199353
Client Ref. No.: 19/MED/901 US Reg

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

In re Application of:	Confirmation No.: 1256
Yuk-Ming Dennis Lo	Examiner: Laura B. Goddard
Application No.: 16/995,607	Technology Center/Art Unit: 1642
Filed: August 17, 2020	
For: DETERMINATION OF BASE MODIFICATIONS OF NUCLEIC ACIDS	SUPPLEMENTAL INFORMATION DISCLOSURE STATEMENT
Customer No.: 20350	

Commissioner for Patents
P.O. Box 1450
Alexandria, VA 22313-1450

Commissioner:

The references cited on attached form PTO/SB/08A are being called to the attention of the Examiner. In accordance with the provisions of 37 CFR §1.98(a)(2), copies of any cited U.S. Patent and U.S. Patent Application Publications are not provided. Copies of the remaining cited references are provided.

It is respectfully requested that the cited references be expressly considered during the prosecution of this application, and the references be made of record therein and appear among the "references cited" on any patent to issue therefrom.

As provided for by 37 CFR §1.97(g) and (h), no inference should be made that the information and references cited are prior art merely because they are in this statement and no

representation is being made that a search has been conducted or that this statement encompasses all the possible relevant information.

This IDS is being filed after an action closing prosecution, but before the payment of the Issue Fee.

I hereby certify under 37 CFR §1.97(e)(1) that each item of information contained in this Information Disclosure Statement was first cited in any communication from a foreign patent office in a counterpart foreign application not more than three months prior to the filing of this statement.

The Commissioner is authorized to charge the fee set forth in §1.17(p) to the firm's credit card for consideration of this paper. The Commissioner is authorized to charge any additional fee due to Deposit Account No. 20-1430.

Respectfully submitted,

/David D. Hsu/

David D. Hsu
Registration No. 71659

KILPATRICK TOWNSEND & STOCKTON LLP

Attachment

74806894V.1

Electronic Patent Application Fee Transmittal

Application Number:	16995607
Filing Date:	17-Aug-2020
Title of Invention:	DETERMINATION OF BASE MODIFICATIONS OF NUCLEIC ACIDS
First Named Inventor/Applicant Name:	Yuk-Ming Dennis Lo
Filer:	David Dah-Wei Hsu/Heather Orosco
Attorney Docket Number:	080015-028410US-1199353

Filed as Small Entity

Filing Fees for Utility under 35 USC 111(a)

Description	Fee Code	Quantity	Amount	Sub-Total in USD($)
Basic Filing:				
Pages:				
Claims:				
Miscellaneous-Filing:				
Petition:				
Patent-Appeals-and-Interference:				
Post-Allowance-and-Post-Issuance:				
Extension-of-Time:				

Description	Fee Code	Quantity	Amount	Sub-Total in USD($)
Miscellaneous:				
SUBMISSION- INFORMATION DISCLOSURE STMT	2806	1	130	130
		Total in USD ($)		**130**

Electronic Acknowledgement Receipt

EFS ID:	42778399
Application Number:	16995607
International Application Number:	
Confirmation Number:	1256
Title of Invention:	DETERMINATION OF BASE MODIFICATIONS OF NUCLEIC ACIDS
First Named Inventor/Applicant Name:	Yuk-Ming Dennis Lo
Customer Number:	20350
Filer:	David Dah-Wei Hsu/Heather Orosco
Filer Authorized By:	David Dah-Wei Hsu
Attorney Docket Number:	080015-028410US-1199353
Receipt Date:	20-MAY-2021
Filing Date:	17-AUG-2020
Time Stamp:	15:53:24
Application Type:	Utility under 35 USC 111(a)

Payment information:

Submitted with Payment	yes
Payment Type	CARD
Payment was successfully received in RAM	$130
RAM confirmation Number	E20215JF54152125
Deposit Account	201430
Authorized User	Heather Orosco

The Director of the USPTO is hereby authorized to charge indicated fees and credit any overpayment as follows:

37 CFR 1.16 (National application filing, search, and examination fees)

37 CFR 1.17 (Patent application and reexamination processing fees)

37 CFR 1.19 (Document supply fees)

37 CFR 1.21 (Miscellaneous fees and charges)

File Listing:

Document Number	Document Description	File Name	File Size(Bytes)/ Message Digest	Multi Part /.zip	Pages (if appl.)
1		1199353_Supplemental_IDS.pdf	218877 79cc47a43ca0cf8d9888f8f07ca07278ba693e8f	yes	4
	Multipart Description/PDF files in .zip description				
	Document Description		**Start**	**End**	
	Information Disclosure Statement (IDS) Form (SB08)		3	4	
	Transmittal Letter		1	2	
Warnings:					
Information:					
2	Non Patent Literature	REF_Examination_Report_for_GB2102808_9_dated_2021-04-12.pdf	302052 33ff994e87575a8090d413944e2083f640dbbc79	no	6
Warnings:					
Information:					
3	Non Patent Literature	REF_Combined_Search_and_Exam_Report_for_GB2103010_1_dated_2021-04-15.pdf	170740 13caea651971598ce47d903648ee7fd08fdd7254	no	4
Warnings:					
Information:					
4	Non Patent Literature	REF_He_2017_Tet1_promotes_leukemic_growth_of_AML1-ETO.pdf	1015598 a7f9d7b2c6313b39d7e0e3accb333fa6ab041ad0	no	2
Warnings:					
Information:					
5	Fee Worksheet (SB06)	fee-info.pdf	30416 89bb55676ba2aced7a3c4c5aa6e0f011c520fab6	no	2
Warnings:					
Information:					

Total Files Size (in bytes):	1737683

This Acknowledgement Receipt evidences receipt on the noted date by the USPTO of the indicated documents, characterized by the applicant, and including page counts, where applicable. It serves as evidence of receipt similar to a Post Card, as described in MPEP 503.

New Applications Under 35 U.S.C. 111

If a new application is being filed and the application includes the necessary components for a filing date (see 37 CFR 1.53(b)-(d) and MPEP 506), a Filing Receipt (37 CFR 1.54) will be issued in due course and the date shown on this Acknowledgement Receipt will establish the filing date of the application.

National Stage of an International Application under 35 U.S.C. 371

If a timely submission to enter the national stage of an international application is compliant with the conditions of 35 U.S.C. 371 and other applicable requirements a Form PCT/DO/EO/903 indicating acceptance of the application as a national stage submission under 35 U.S.C. 371 will be issued in addition to the Filing Receipt, in due course.

New International Application Filed with the USPTO as a Receiving Office

If a new international application is being filed and the international application includes the necessary components for an international filing date (see PCT Article 11 and MPEP 1810), a Notification of the International Application Number and of the International Filing Date (Form PCT/RO/105) will be issued in due course, subject to prescriptions concerning national security, and the date shown on this Acknowledgement Receipt will establish the international filing date of the application.

UNITED STATES PATENT AND TRADEMARK OFFICE

UNITED STATES DEPARTMENT OF COMMERCE
United States Patent and Trademark Office
Address: COMMISSIONER FOR PATENTS
P.O. Box 1450
Alexandria, Virginia 22313-1450
www.uspto.gov

APPLICATION NO.	FILING DATE	FIRST NAMED INVENTOR	ATTORNEY DOCKET NO.	CONFIRMATION NO.
16/995,607	08/17/2020	Yuk-Ming Dennis Lo	080015-028410US-1199353	1256

20350 7590 04/26/2021

Kilpatrick Townsend & Stockton LLP - West Coast
Mailstop: IP Docketing - 22
1100 Peachtree Street
Suite 2800
Atlanta, GA 30309

EXAMINER
GODDARD, LAURA B

ART UNIT	PAPER NUMBER
1642	

NOTIFICATION DATE	DELIVERY MODE
04/26/2021	ELECTRONIC

Please find below and/or attached an Office communication concerning this application or proceeding.

The time period for reply, if any, is set in the attached communication.

Notice of the Office communication was sent electronically on above-indicated "Notification Date" to the following e-mail address(es):

KTSDocketing2@kilpatrick.foundationip.com
ipefiling@kilpatricktownsend.com

Notice of Allowability	Application No. 16/995,607	Applicant(s) Lo et al.	
	Examiner LAURA B GODDARD	Art Unit 1642	AIA (FITF) Status Yes

-- The MAILING DATE of this communication appears on the cover sheet with the correspondence address--

All claims being allowable, PROSECUTION ON THE MERITS IS (OR REMAINS) CLOSED in this application. If not included herewith (or previously mailed), a Notice of Allowance (PTOL-85) or other appropriate communication will be mailed in due course. **THIS NOTICE OF ALLOWABILITY IS NOT A GRANT OF PATENT RIGHTS.** This application is subject to withdrawal from issue at the initiative of the Office or upon petition by the applicant. See 37 CFR 1.313 and MPEP 1308.

1. ☑ This communication is responsive to 4/15/2021.
   ☐ A declaration(s)/affidavit(s) under **37 CFR 1.130(b)** was/were filed on _____.

2. ☐ An election was made by the applicant in response to a restriction requirement set forth during the interview on _____; the restriction requirement and election have been incorporated into this action.

3. ☑ The allowed claim(s) is/are 1-17 and 70-71 . As a result of the allowed claim(s), you may be eligible to benefit from the **Patent Prosecution Highway** program at a participating intellectual property office for the corresponding application. For more information , please see **http://www.uspto.gov/patents/init_events/pph/index.jsp** or send an inquiry to **PPHfeedback@uspto.gov.**

4. ☐ Acknowledgment is made of a claim for foreign priority under 35 U.S.C. § 119(a)-(d) or (f).

   Certified copies:

   a) ☐ All b) ☐ Some *c) ☐ None of the:

   1. ☐ Certified copies of the priority documents have been received.
   2. ☐ Certified copies of the priority documents have been received in Application No. _____ .
   3. ☐ Copies of the certified copies of the priority documents have been received in this national stage application from the International Bureau (PCT Rule 17.2(a)).

   * Certified copies not received: _____ .

Applicant has THREE MONTHS FROM THE "MAILING DATE" of this communication to file a reply complying with the requirements noted below. Failure to timely comply will result in ABANDONMENT of this application.
THIS THREE-MONTH PERIOD IS NOT EXTENDABLE.

5. ☐ CORRECTED DRAWINGS (as "replacement sheets") must be submitted.
   ☐ including changes required by the attached Examiner's Amendment / Comment or in the Office action of Paper No./Mail Date _____ .

   Identifying indicia such as the application number (see 37 CFR 1.84(c)) should be written on the drawings in the front (not the back) of each sheet. Replacement sheet(s) should be labeled as such in the header according to 37 CFR 1.121(d).

6. ☐ DEPOSIT OF and/or INFORMATION about the deposit of BIOLOGICAL MATERIAL must be submitted. Note the attached Examiner's comment regarding REQUIREMENT FOR THE DEPOSIT OF BIOLOGICAL MATERIAL.

Attachment(s)

1. ☐ Notice of References Cited (PTO-892)
2. ☑ Information Disclosure Statements (PTO/SB/08), Paper No./Mail Date _____.
3. ☐ Examiner's Comment Regarding Requirement for Deposit of Biological Material _____.
4. ☐ Interview Summary (PTO-413), Paper No./Mail Date. _____.
5. ☐ Examiner's Amendment/Comment
6. ☐ Examiner's Statement of Reasons for Allowance
7. ☐ Other _____.

/Laura B Goddard/
Primary Examiner, Art Unit 1642

INFORMATION DISCLOSURE STATEMENT BY APPLICANT (Not for submission under 37 CFR 1.99)		
	Application Number	16/995,607
	Filing Date	August 17, 2020
	First Named Inventor	Yuk-Ming Dennis Lo
	Art Unit	1642
	Examiner Name	Laura B. Goddard
	Attorney Docket Number	080015-028410US-1199353

U.S. PATENTS

Examiner Initial*	Cite No	Patent Number	Kind Code [1]	Issue Date	Name of Patentee or Applicant of cited Document	Pages, Columns, Lines, Where Relevant Passages or Relevant Figures Appear

U.S. PATENT APPLICATION PUBLICATIONS

Examiner Initial*	Cite No	Publication Number	Kind Code [1]	Publication Date	Name of Patentee or Applicant of cited Document	Pages, Columns, Lines, Where Relevant Passages or Relevant Figures Appear
	1.	2016/0017419	A1	01-21-2016	Chiu et al.	

FOREIGN PATENT DOCUMENTS

Examiner Initial*	Cite No	Foreign Document Number[3]	Country Code[2]	Kind Code[4]	Publication Date	Name of Patentee or Applicant of cited Document	Pages, Columns, Lines where Relevant Passages or Relevant Figures Appear	T[5]
	2.	2017012544	WO	A1	01-26-2017	The Chinese University of Hong Kong		

NON-PATENT LITERATURE DOCUMENTS

Examiner Initials*	Cite No	Include name of the author (in CAPITAL LETTERS), title of the article (when appropriate), title of the item (book, magazine, journal, serial, symposium, catalog, etc.), date, page(s), volume-issue number(s), publisher, city and/or country where published.	T[6]
	3.	Combined Search and Examination Report under Sections 17 & 18(3) dated March 29, 2021 in GB Patent Application No. 2103003.6. 5 pages.	

EXAMINER SIGNATURE

Examiner Signature	/Laura B Goddard/	Date Considered	04/22/2021	

*EXAMINER: Initial if reference considered, whether or not citation is in conformance with MPEP 609. Draw line through a citation if not in conformance and not considered. Include copy of this form with next communication to applicant.

ALL REFERENCES CONSIDERED EXCEPT WHERE LINED THROUGH. /LBG/

INFORMATION DISCLOSURE STATEMENT BY APPLICANT (Not for submission under 37 CFR 1.99)		
	Application Number	16/995,607
	Filing Date	August 17, 2020
	First Named Inventor	Yuk-Ming Dennis Lo
	Art Unit	1642
	Examiner Name	Laura B. Goddard
	Attorney Docket Number	080015-028410US-1199353

[1] See Kind Codes of USPTO Patent Documents at www.USPTO.GOV or MPEP 901.04. [2] Enter office that issued the document, by the two-letter code (WIPO Standard ST.3). [3] For Japanese patent documents, the indication of the year of the reign of the Emperor must precede the serial number of the patent document.
[4] Kind of document by the appropriate symbols as indicated on the document under WIPO Standard ST.16 if possible. [5] Applicant is to place a check mark here if English language translation is attached.

ALL REFERENCES CONSIDERED EXCEPT WHERE LINED THROUGH. /LBG/

INFORMATION DISCLOSURE STATEMENT BY APPLICANT (Not for submission under 37 CFR 1.99)		
	Application Number	16/995,607
	Filing Date	August 17, 2020
	First Named Inventor	Yuk-Ming Dennis Lo
	Art Unit	1642
	Examiner Name	Laura B. Goddard
	Attorney Docket Number	080015-028410US-1199353

U.S. PATENTS

Examiner Initial*	Cite No	Patent Number	Kind Code [1]	Issue Date	Name of Patentee or Applicant of cited Document	Pages, Columns, Lines, Where Relevant Passages or Relevant Figures Appear

U.S. PATENT APPLICATION PUBLICATIONS

Examiner Initial*	Cite No	Publication Number	Kind Code [1]	Publication Date	Name of Patentee or Applicant of cited Document	Pages, Columns, Lines, Where Relevant Passages or Relevant Figures Appear
	1.	2016/0017419	A1	01-21-2016	Chiu et al.	

FOREIGN PATENT DOCUMENTS

Examiner Initial*	Cite No	Foreign Document Number[3]	Country Code[2]	Kind Code[4]	Publication Date	Name of Patentee or Applicant of cited Document	Pages, Columns, Lines where Relevant Passages or Relevant Figures Appear	T[5]
	2.	2017012544	WO	A1	01-26-2017	The Chinese University of Hong Kong		

NON-PATENT LITERATURE DOCUMENTS

Examiner Initials*	Cite No	Include name of the author (in CAPITAL LETTERS), title of the article (when appropriate), title of the item (book, magazine, journal, serial, symposium, catalog, etc.), date, page(s), volume-issue number(s), publisher, city and/or country where published.	T[5]
	3.	Combined Search and Examination Report under Sections 17 & 18(3) dated March 29, 2021 in GB Patent Application No. 2103003.6. 5 pages.	

EXAMINER SIGNATURE

Examiner Signature		Date Considered	

*EXAMINER: Initial if reference considered, whether or not citation is in conformance with MPEP 609. Draw line through a citation if not in conformance and not considered. Include copy of this form with next communication to applicant.

INFORMATION DISCLOSURE STATEMENT BY APPLICANT (Not for submission under 37 CFR 1.99)		
	Application Number	16/995,607
	Filing Date	August 17, 2020
	First Named Inventor	Yuk-Ming Dennis Lo
	Art Unit	1642
	Examiner Name	Laura B. Goddard
	Attorney Docket Number	080015-028410US-1199353

[1] See Kind Codes of USPTO Patent Documents at www.USPTO.GOV or MPEP 901.04. [2] Enter office that issued the document, by the two-letter code (WIPO Standard ST.3). [3] For Japanese patent documents, the indication of the year of the reign of the Emperor must precede the serial number of the patent document.
[4] Kind of document by the appropriate symbols as indicated on the document under WIPO Standard ST.16 if possible. [5] Applicant is to place a check mark here if English language translation is attached.

(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
26 January 2017 (26.01.2017)

WIPO | PCT

(10) International Publication Number
WO 2017/012544 A1

(51) **International Patent Classification**:
C12Q 1/68 (2006.01)

(21) **International Application Number**:
PCT/CN2016/090642

(22) **International Filing Date**:
20 July 2016 (20.07.2016)

(25) **Filing Language**: English

(26) **Publication Language**: English

(30) **Priority Data**:
62/194,702 20 July 2015 (20.07.2015) US

(71) **Applicant**: **THE CHINESE UNIVERSITY OF HONG KONG** [CN/CN]; Office of Research and Knowledge Transfer Services, Room 301, Pi Ch'iu Building, Shatin, New Territories, Hong Kong 999077 (CN).

(72) **Inventors**: **LO, Yuk-Ming Dennis**; 4th Floor, 7 King Tak Street, Homantim, Kowloon, Hong Kong 999077 (CN). **CHAN, Kwan Chee**; Flat C, 15th Floor, Block 5, Peak One, Mei Tin Road, Shatin, New Territories, Hong Kong 999077 (CN). **CHIU, Rossa Wai Kwun**; House 31, Double Haven, 52 Ma Lok Path, Shatin, New Territories, Hong Kong 999077 (CN). **JIANG, Peiyong**; Flat 7 on 1st Floor of Block B of Kwong Lam Court, Nos. 62-66 (No.62) Siu Lek Yuen Road, Shatin, New Territories, Hong Kong 999077 (CN). **SUN, Kun**; Rm 711, Block A, Staff Quarters, Prince of Wales Hospital, 30-32 Ngan Shing Street, Shatin, New Territories, Hong Kong (CN).

(74) **Agent**: **INSIGHT INTELLECTUAL PROPERTY LIMITED**; 19 A, Tower A, InDo Building, No. 48A Zhichun Road, Haidian District, Beijing 100098 (CN).

(81) **Designated States** *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JP, KE, KG, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(84) **Designated States** *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

Published:

— *with international search report (Art. 21(3))*

WO 2017/012544 A1

(54) **Title**: METHYLATION PATTERN ANALYSIS OF HAPLOTYPES IN TISSUES IN DNA MIXTURE

(57) **Abstract**: Systems, apparatuses, and method are provided for determining the contributions of different tissues to a biological sample that includes a mixture of cell-free DNA molecules from various tissues types, e.g., as occurs in plasma or serum and other body fluids. Embodiments can analyze the methylation patterns of the DNA mixture (e.g., methylation levels at particular loci) for a particular haplotype and determine fractional contributions of various tissue types to the DNA mixture, e.g., of fetal tissue types or tissue types of specific organs that might have a tumor. Such fractional contributions determined for a haplotype can be used in a variety of ways.

METHYLATION PATTERN ANALYSIS OF HAPLOTYPES IN TISSUES IN DNA MIXTURE

CROSS-REFERENCES TO RELATED APPLICATIONS

[0001] The present application claims priority to U.S. Provisional Application No. 62/194,702, entitled "Methylation Pattern Analysis Of Haplotypes In Tissues In A DNA Mixture" filed July 20, 2015, the entire contents of which is herein incorporated by reference for all purposes.

BACKGROUND

[0002] It has previously been demonstrated that through the analysis of plasma DNA of a pregnant woman carrying a fetus, the maternal haplotypes inherited by the fetus can be deduced using the process of relative haplotype dosage analysis (RHDO) (Lo et al. Sci Transl Med 2010; 2: 61ra91 and US Patent 8,467,976). The haplotype information for the pregnant woman can be used. The haplotype information can be obtained using family analysis or a method for the direct analysis of the haplotype (e.g. Fan et al. Nat Biotechnol 2011; 29: 51-57; Snyder et al. Nat Rev Genet 2015; 16: 344-358). SNPs that are heterozygous in the mother but homozygous in the father can be used for the RHDO analysis.

[0003] Such a use of specific SNPs can limit the loci that can be used, and therefore limit the amount of data and accuracy. Such use of specific SNPs may also limit the clinical utility of the method as DNA samples from additional family members may not be available, and methods for the direct analysis of haplotypes would add costs to the analysis.

BRIEF SUMMARY

[0004] Embodiments are described for determining the contributions of different tissues to a biological sample that includes a mixture of cell-free DNA molecules from various tissues types, e.g., as occurs in plasma or serum and other body fluids. Embodiments can analyze the methylation patterns of the DNA mixture (e.g., methylation levels at particular loci) for a particular haplotype and determine fractional contributions of various tissue types to the DNA

mixture, e.g., of fetal tissue types or tissue types of specific organs that might have a tumor. Such fractional contributions determined for a haplotype can be used in a variety of ways.

[0005] In some embodiments, two fractional contributions of a tissue type can be determined using methylation levels of two sets of cell-free DNA molecules from a maternal sample, each set being for a different one of two parental haplotypes of a parent of a fetus, for a chromosomal region being analyzed. In various implementations, the maternal sample can be the plasma or serum sample from a woman pregnant with one or more fetuses. The two fractional contributions can be used to identify a portion of the fetal genome. For example, a separation value between the two fractional contributions of fetal tissue can indicate the fetal genotype at a locus and can indicate which of the two parental haplotypes is inherited by the fetus. For instance, the higher fractional contribution can indicate the inherited haplotype, and both can be inherited if the separation value is less than a threshold; both haplotypes could be inherited when both parents share the haplotype (or allele for a genotype) for the region being analyzed.

[0006] In some embodiments, only one fractional contribution of fetal tissue is determined for one haplotype. When the one fractional contribution exceeds a reference value (e.g., as determined from other samples), the fetus can be determined to have inherited the one haplotype for the region being analyzed.

[0007] In some embodiments, two methylation levels can be determined for two sets of cell-free DNA molecules from a maternal sample, each set being for a different one of two parental haplotypes of a parent of a fetus, as part of identifying a portion of the fetal genome. The two methylation levels can be compared to each other to identify which haplotype is inherited by the fetus, e.g., by which methylation level is lower. For example, a fetus contributes cell-free DNA molecules that are hypomethylated, and a measurement of a lower methylation level of one haplotype indicates that the one haplotype is inherited by the fetus.

[0008] In some embodiments, a sequence imbalance can be detected for a target chromosomal region of a fetus using a mixture of cell-free DNA molecules from a plurality of tissues types. Target heterozygous loci can be identified for the target chromosomal region having a first target haplotype and a second target haplotype having different alleles. A first target fractional contribution of the fetal tissue type in the mixture can be determined using methylation levels at the target heterozygous loci, where the methylation levels are determined using a target set of

cell-free DNA molecules located at (i.e., covering) the loci of the first haplotype. Similarly, a first reference fractional contribution of the fetal tissue type can be determined. A separation value of the first target fractional contribution and the first reference fractional contribution can be compared to a threshold value to determine whether the fetus has a sequence imbalance. If the two fractional contributions are significantly different, then a sequence imbalance can be determined. The specific threshold being used can depend on the specific sequence imbalance (e.g., an amplification or a deletion) being tested.

[0009] In some embodiments, a fractional contribution of a first haplotype in a first tissue type can be used to determine whether the first tissue type has a disease state. The first haplotype can have a signature specific to healthy cells or to abnormal cells. Thus, the first haplotype can be not present in healthy cells of the organism, or present in healthy cells of the organism and not present in abnormal cells that may be in the mixture. A separation value between the first fractional contribution and a reference fractional contribution can be compared to a threshold value to determine a classification of whether the first tissue type has a disease state.

[0010] In some embodiments, the tissue origin of a copy number aberration can be determined using methylation deconvolution. A first chromosomal region can be identified as exhibiting a copy number aberration. The haplotypes in the first chromosomal region can be determined. For each of M tissue types, a corresponding separation value between the corresponding first fractional contribution and the corresponding second fractional contribution can be determined. The tissue type with the highest separation value can be identified as the tissue of origin.

[0011] Other embodiments are directed to systems and computer readable media associated with methods described herein.

[0012] A better understanding of the nature and advantages of embodiments of the present invention may be gained with reference to the following detailed description and the accompanying drawings.

BRIEF DESCRIPTION OF THE DRAWINGS

[0013] FIG. 1 is a flowchart illustrating a method of analyzing a DNA mixture of cell-free DNA molecules to determine fractional contributions from various tissue types from methylation levels according to embodiments of the present invention.

[0014] FIG. 2 shows a schematic diagram showing several potential applications of DNA methylation deconvolution (e.g., using plasma) and its applications according to embodiments of the present invention.

[0015] FIG. 3A shows a graph of percentage contributions of different organs to plasma DNA for 15 pregnant women according to embodiments of the present invention. FIG. 3B shows a plot 350 of a correlation between the plasma DNA fractions contributed by the placenta deduced from plasma DNA methylation deconvolution and the fetal DNA fractions deduced using fetal-specific SNP alleles according to embodiments of the present invention.

[0016] FIG. 4 shows a table of percentage contributions determined from a plasma DNA tissue mapping analysis among pregnant women according to embodiments of the present invention.

[0017] FIG. 5 shows plots of percentage contributions of organs other than the placenta by plasma DNA tissue mapping and fetal DNA fractions based on fetal-specific SNP alleles according to embodiments of the present invention.

[0018] FIG. 6 shows a table of percentage contributions from plasma DNA tissue mapping analysis among the non-pregnant healthy control subjects according to embodiments of the present invention.

[0019] FIG. 7 shows a table of the estimated contributions of different organs to the plasma DNA for 11 pregnant women and 4 non-pregnant healthy subjects using the first set of markers (with high organ specificity) according to embodiments of the present invention.

[0020] FIG. 8 shows a table of the estimated contributions of different organs to the plasma DNA for 11 pregnant women and 4 non-pregnant healthy subjects using the second set of markers (with low organ specificity) according to embodiments of the present invention.

[0021] FIG. 9A is a plot showing the correlation between the estimated fetal DNA fraction (contribution from the placenta) and the fetal DNA fraction determined by counting the fetal-specific alleles in the maternal plasma samples.

[0022] FIG. 9B is a plot showing absolute difference between the estimation from methylation markers and fetal DNA fraction determined by fetal-specific alleles counting.

[0023] FIG. 10A is a graph showing placental contribution to plasma DNA deduced using markers with different selection criteria according to embodiments of the present invention. FIG. 10B is a graph showing the accuracy of plasma DNA deconvolution using markers with low variability (category i) and high variability (category ii) in the same type of tissue.

[0024] FIG. 11A shows a first scenario where the fetus has inherited the M allele from the mother and has a genotype of MN at a particular locus according to embodiments of the present invention. FIG. 11B shows a second scenario where the fetus has inherited the N allele from the mother and has the genotype of NN at a particular locus according to embodiments of the present invention.

[0025] FIG. 12A shows a determination of a maternal haplotype inherited by a fetus using methylation deconvolution according to embodiments of the present invention. FIG. 12B shows an illustration of the paternal haplotype methylation analysis according to embodiments of the present invention.

[0026] FIG. 13 is a flowchart illustrating a method 1300 of determining a portion of a fetal genome from a maternal sample using methylation deconvolution according to embodiments of the present invention.

[0027] FIG. 14 is a flowchart illustrating a method 1400 of determining a portion of a fetal genome from a maternal sample using methylation levels according to embodiments of the present invention.

[0028] FIG. 15 shows chromosomal aneuploidy detection based on haplotype deconvolution for maternal haplotypes according to embodiments of the present invention.

[0029] FIG. 16 shows chromosomal aneuploidy detection based on haplotype deconvolution for paternal haplotypes according to embodiments of the present invention.

[0030] FIG. 17 is a flowchart of a method 1700 for detecting a sequence imbalance in a portion of a fetal genome of an unborn fetus of a pregnant female using a biological sample from the pregnant female according to embodiments of the present invention.

[0031] FIG. 18 shows an illustration of haplotype deconvolution for organ transplant monitoring according to embodiments of the present invention.

[0032] FIG. 19 is a flowchart illustrating a method of analyzing a biological sample of an organism to detect whether a first tissue type has a disease state associated with a first haplotype according to embodiments of the present invention.

[0033] FIG. 20 shows a plot of copy number aberrations detected in the plasma of an HCC patient according to embodiments of the present invention.

[0034] FIG. 21 is a flowchart illustrating a method of analyzing a biological sample of an organism to identify an origin of a chromosomal aberration according to embodiments of the present invention.

[0035] FIG. 22 shows a block diagram of an example computer system 10 usable with system and methods according to embodiments of the present invention.

TERMS

[0036] A "methylome" provides a measure of an amount of DNA methylation at a plurality of sites or loci in a genome. The methylome may correspond to all of the genome, a substantial part of the genome, or relatively small portion(s) of the genome. Examples of methylomes of interest are the methylomes of organs (e.g. methylomes of brain cells, bones, the lungs, the heart, the muscles and the kidneys, etc.) that can contribute DNA into a bodily fluid (e.g. plasma, serum, sweat, saliva, urine, genital secretions, semen, stools fluid, diarrheal fluid, cerebrospinal fluid, secretions of the gastrointestinal tract, ascitic fluid, pleural fluid, intraocular fluid, fluid from a hydrocele (e.g. of the testis), fluid from a cyst, pancreatic secretions, intestinal secretions, sputum, tears, aspiration fluids from breast and thyroid, etc.). The organs may be transplanted organs. The methylome of a fetus is another example.

[0037] A "plasma methylome" is a methylome determined from the plasma or serum of an animal (e.g., a human). The plasma methylome is an example of a cell-free methylome since

plasma and serum include cell-free DNA. The plasma methylome is also an example of a mixed methylome since it is a mixture of fetal/maternal methylome or tumor/patient methylome or DNA derived from different tissues or organs or donor/recipient methylome in the context or organ transplantation.

[0038] A "site" (also called a "genomic site") corresponds to a single site, which may be a single base position or a group of correlated base positions, e.g., a CpG site or larger group of correlated base positions. A "locus" may correspond to a region that includes multiple sites. A locus can include just one site, which would make the locus equivalent to a site in that context.

[0039] The "methylation index" for each genomic site (e.g., a CpG site) can refer to the proportion of DNA fragments (e.g., as determined from sequence reads or probes) showing methylation at the site over the total number of reads covering that site. A "read" can correspond to information (e.g., methylation status at a site) obtained from a DNA fragment. A read can be obtained using reagents (e.g. primers or probes) that preferentially hybridize to DNA fragments of a particular methylation status. Typically, such reagents are applied after treatment with a process that differentially modifies or differentially recognizes DNA molecules depending of their methylation status, e.g. bisulfite conversion, or methylation-sensitive restriction enzyme, or methylation binding proteins, or anti-methylcytosine antibodies, or single molecule sequencing techniques that recognize methylcytosines and hydroxymethylcytosines.

[0040] The "methylation density" of a region can refer to the number of reads at sites within the region showing methylation divided by the total number of reads covering the sites in the region. The sites may have specific characteristics, e.g., being CpG sites. Thus, the "CpG methylation density" of a region can refer to the number of reads showing CpG methylation divided by the total number of reads covering CpG sites in the region (e.g., a particular CpG site, CpG sites within a CpG island, or a larger region). For example, the methylation density for each 100-kb bin in the human genome can be determined from the total number of cytosines not converted after bisulfite treatment (which corresponds to methylated cytosine) at CpG sites as a proportion of all CpG sites covered by sequence reads mapped to the 100-kb region. This analysis can also be performed for other bin sizes, e.g. 500 bp, 5 kb, 10 kb, 50-kb or 1-Mb, etc. A region could be the entire genome or a chromosome or part of a chromosome (e.g. a chromosomal arm). The methylation index of a CpG site is the same as the methylation density for a region when the

region only includes that CpG site. The "proportion of methylated cytosines" can refer the number of cytosine sites, "C's", that are shown to be methylated (for example unconverted after bisulfite conversion) over the total number of analyzed cytosine residues, i.e. including cytosines outside of the CpG context, in the region. The methylation index, methylation density and proportion of methylated cytosines are examples of "methylation levels." Apart from bisulfite conversion, other processes known to those skilled in the art can be used to interrogate the methylation status of DNA molecules, including, but not limited to enzymes sensitive to the methylation status (e.g. methylation-sensitive restriction enzymes), methylation binding proteins, single molecule sequencing using a platform sensitive to the methylation status (e.g. nanopore sequencing (Schreiber et al. Proc Natl Acad Sci 2013; 110: 18910-18915) and by the Pacific Biosciences single molecule real time analysis (Flusberg et al. Nat Methods 2010; 7: 461-465)).

[0041] A "methylation profile" (also called methylation status) includes information related to DNA methylation for a region. Information related to DNA methylation can include, but not limited to, a methylation index of a CpG site, a methylation density of CpG sites in a region, a distribution of CpG sites over a contiguous region, a pattern or level of methylation for each individual CpG site within a region that contains more than one CpG site, and non-CpG methylation. A methylation profile of a substantial part of the genome can be considered equivalent to the methylome. "DNA methylation" in mammalian genomes typically refers to the addition of a methyl group to the 5' carbon of cytosine residues (i.e. 5-methylcytosines) among CpG dinucleotides. DNA methylation may occur in cytosines in other contexts, for example CHG and CHH, where H is adenine, cytosine or thymine. Cytosine methylation may also be in the form of 5-hydroxymethylcytosine. Non-cytosine methylation, such as N^6-methyladenine, has also been reported.

[0042] "Methylation-aware sequencing" refers to any sequencing method that allows one to ascertain the methylation status of a DNA molecule during a sequencing process, including, but not limited to bisulfite sequencing, or sequencing preceded by methylation-sensitive restriction enzyme digestion, immunoprecipitation using anti-methylcytosine antibody or methylation binding protein, or single molecule sequencing that allows elucidation of the methylation status.

[0043] A "tissue" corresponds to a group of cells that group together as a functional unit. More than one type of cells can be found in a single tissue. Different types of tissue may consist of

different types of cells (e.g., hepatocytes, alveolar cells or blood cells), but also may correspond to tissue from different organisms (mother vs. fetus) or to healthy cells vs. tumor cells. "Reference tissues" correspond to tissues used to determine tissue-specific methylation levels. Multiple samples of a same tissue type from different individuals may be used to determine a tissue-specific methylation level for that tissue type.

[0044] A "biological sample" refers to any sample that is taken from a subject (*e.g.*, a human, such as a pregnant woman, a person with cancer, or a person suspected of having cancer, an organ transplant recipient or a subject suspected of having a disease process involving an organ (e.g., the heart in myocardial infarction, or the brain in stroke, or the hematopoietic system in anemia) and contains one or more nucleic acid molecule(s) of interest. The biological sample can be a bodily fluid, such as blood, plasma, serum, urine, vaginal fluid, fluid from a hydrocele (e.g. of the testis), or vaginal flushing fluids, pleural fluid, ascitic fluid, cerebrospinal fluid, saliva, sweat, tears, sputum, bronchoalveolar lavage fluid, etc. Stool samples can also be used. In various embodiments, the majority of DNA in a biological sample that has been enriched for cell-free DNA (e.g., a plasma sample obtained via a centrifugation protocol) can be cell-free (as opposed to cells), e.g., greater than 50%, 60%, 70%, 80%, 90%, 95%, or 99%. The centrifugation protocol can include 3,000 g x 10 minutes, obtaining the fluid part, and recentrifuging at 30,000 g for another 10 minutes to remove residual cells.

[0045] The term "level of cancer" can refer to whether cancer exists (i.e., presence or absence), a stage of a cancer, a size of tumor, whether there is metastasis, the total tumor burden of the body, and/or other measure of a severity of a cancer (e.g. recurrence of cancer). The level of cancer could be a number or other indicia, such as symbols, alphabet letters, and colors. The level could be zero. The level of cancer also includes premalignant or precancerous conditions (states) associated with mutations or a number of mutations. The level of cancer can be used in various ways. For example, screening can check if cancer is present in someone who is not known previously to have cancer. Assessment can investigate someone who has been diagnosed with cancer to monitor the progress of cancer over time, study the effectiveness of therapies or to determine the prognosis. In one embodiment, the prognosis can be expressed as the chance of a patient dying of cancer, or the chance of the cancer progressing after a specific duration or time, or

the chance of cancer metastasizing. Detection can mean 'screening' or can mean checking if someone, with suggestive features of cancer (e.g. symptoms or other positive tests), has cancer.

[0046] The term "sequence imbalance" of a chromosomal region can refer to any significant deviation in an amount of cell-free DNA molecules from the chromosomal region relative to an expected value, if the organism was healthy. For example, a chromosomal region may exhibit an amplification or a deletion in a certain tissue, thereby resulting in a sequence imbalance for the chromosomal region in a DNA mixture containing DNA from the tissue, mixed with DNA from other tissues. As examples, the expected value can be obtained from another sample or from another chromosomal region that is assumed to be normal (e.g., an amount representative of two copies for a diploid organism). A chromosomal region can be composed of multiple disjoint subregions.

[0047] A "type" for a genomic locus (marker) corresponds to specific attributes for a locus across tissue types. The description primarily refers to type I loci and type II loci, whose properties are provided in detail below. A locus of a given type can have specific statistical variation in methylation levels across tissue types. A "category" for a genomic locus (marker) corresponds to specific variation in methylation levels for a locus across different individuals for a same tissue type. A set of genomic loci (markers) can be composed of any number of loci of various types and/or categories. Thus, a set of loci corresponds to loci selected for a particular measurement and does not connote any particular properties of the loci in the set.

[0048] A "separation value" corresponds to a difference or a ratio involving two values, e.g., two fractional contributions or two methylation levels. The separation value could be a simple difference or ratio. The separation value can include other factors, e.g., multiplicative factors. As other examples, a difference or ratio of functions of the values can be used, e.g., a difference or ratio of the natural logarithms (ln) of the two values. A separation value can include a difference and a ratio.

[0049] The term "*classification*" as used herein refers to any number(s) or other characters(s) that are associated with a particular property of a sample. For example, a "+" symbol (or the word "positive") could signify that a sample is classified as having deletions or amplifications. The classification can be binary (e.g., positive or negative) or have more levels of classification (e.g., a scale from 1 to 10 or 0 to 1). The term "*cutoff*" and "*threshold*" refer to a predetermined number

used in an operation. A threshold value may be a value above or below which a particular classification applies. Either of these terms can be used in either of these contexts.

DETAILED DESCRIPTION

[0050] Methylation differences among tissue types (e.g., fetal tissues, liver etc.) in a DNA mixture (e.g., plasma) may be used to differentiate properties of haplotypes in a specific tissue type. For example, the methylation levels of two maternal haplotypes in the plasma of a pregnant woman can be used to determine which haplotype is inherited from the mother to the fetus. As another example, the methylation levels of two haplotypes in fetal tissue can be used to detect a sequence imbalance (e.g., an aneuploidy) in the fetus. Other tissue types may also be analyzed, e.g., to detect a disease state in a specific tissue type. The tissue type from which a copy number aberration originates may also be determined.

[0051] Some embodiments can determine percentages of cell-free DNA in plasma (or other DNA mixture) from various tissue types using known methylation levels at certain genomic sites for the specific tissue types. For example, methylation levels at the genomic sites can be measured for a liver sample, and these tissue-specific methylation levels can be used to determine how much cell-free DNA in the mixture is from the liver. Methylation levels can be measured for tissue types that provide substantial contributions to the DNA mixture, so that a predominance (e.g., more than 90%, 95%, or 99%) of the cell-free DNA mixture can be accounted for. Such other samples can include, but not limited to, some or all of the following: lung, colon, small intestines, pancreas, adrenal glands, esophagus, adipose tissues, heart, and brain.

[0052] A deconvolution process can be used to determine fractional contributions (e.g., percentage) for each of the tissue types for which tissue-specific methylation levels are known. In some embodiments, a linear system of equations can be created from the known tissue-specific methylation levels and the mixture methylation levels at the specified genomic sites, and the fractional contributions that best approximate the measured mixture methylation levels can be determined (e.g., using least squares).

[0053] Once the fractional contributions are determined, the fractional contributions can be used for various purposes. For example, differences in fractional contributions of fetal tissue can be used to determine which haplotype is inherited from a parent. The alleles at one or more

heterozygous loci can be determined for each of two parental haplotypes. Cell-free DNA at the one or more heterozygous loci can be used to determine two fractional contributions: one for each haplotype. For example, cell-free DNA molecules having alleles of a first haplotype can be used to determine a first fractional contribution, and cell-free DNA molecules having alleles of a second haplotype can be used to determine a second fractional contribution. The inherited haplotype will correspond to the higher fractional contribution for fetal tissue.

[0054] Further, an inherited haplotype will have lower methylation levels due to the general hypomethylation of fetal cell-free DNA. The methylation levels for the two haplotypes can be compared, and the haplotype with lower methylation level can be identified as being the inherited haplotype.

[0055] As another example, a sequence imbalance can be detected in a target chromosomal region of a fetus. A target fractional contribution of the fetal tissue type in the mixture can be determined for a first haplotype in the target chromosomal region. Similarly, a reference fractional contribution of the fetal tissue type can be determined for a reference chromosomal region. A separation value between the two contributions can be compared to a threshold value to determine whether the fetus has a sequence imbalance (e.g., an aneuploidy).

[0056] As another example, a first haplotype can have a signature specific to healthy cells or to abnormal cells. A separation value between the fractional contribution determined for the first haplotype and a reference fractional contribution can be compared to a threshold value to determine a classification of whether the first tissue type has a disease state. As examples, the first haplotype can be of a transplanted organ or a tumor, or only be in healthy cells and not in a transplanted organ or a tumor. The disease state may be whether the transplanted organ is being rejected, or whether a tumor is increasing in size or metastasized (e.g., after a surgery did not remove all of the tumor).

[0057] As another example, the tissue origin of a copy number aberration can be determined using methylation deconvolution. A first chromosomal region can be identified as exhibiting a copy number aberration. For each of M tissue types, a corresponding separation value between the fractional contributions of the two haplotypes in the first chromosomal region can be determined. The tissue type with the highest separation value can be identified as the tissue of origin.

[0058] Methylation deconvolution is first described, and then the selection of methylation markers and the accuracy of the methylation deconvolution are described. The use of the fractional contributions to determine part of a fetal genome is then described.

I. COMPOSITION OF DNA MIXTURE BY METHYLATION DECONVOLUTION

[0059] Different tissue types can have different levels of methylation for a genomic site. These differences can be used to determine the fractional contributions of DNA from the various tissue types in a mixture. Thus, the composition of a DNA mixture can be determined by a tissue-specific methylation pattern analysis. The examples below discuss methylation densities, but other methylation levels can be used.

A. Single Genomic Site

[0060] The principle of methylation deconvolution can be illustrated using a single methylation genomic site (methylation marker) to determine a composition of a DNA mixture from an organism. Assume that tissue A is completely methylated for the genomic site, i.e. methylation density (MD) of 100% and tissue B is completely unmethylated, i.e. MD of 0%. In this example, methylation density refers to the percentage of cytosine residues with the context of CpG dinucleotides being methylated in the region of interest.

[0061] If the DNA mixture C is composed of tissue A and tissue B and the overall methylation density of the DNA mixture C is 60%, we can deduce the proportional contribution of tissues A and B to the DNA mixture C according to the following formula:

$$MD_C = MD_A \times a + MD_B \times b,$$

where MD_A, MD_B, MD_C represent the MD of tissues A, tissue B and the DNA mixture C, respectively; and a and b are the proportional contributions of tissues A and B to the DNA mixture C. In this particular example, it is assumed that tissues A and B are the only two constituents of the DNA mixture. Therefore, a + b = 100%. Thus, it is calculated that tissues A and B contribute 60% and 40%, respectively, to the DNA mixture.

[0062] The methylation densities in tissue A and tissue B can be obtained from samples of the organism or from samples from other organisms of the same type (e.g., other humans, potentially

of a same subpopulation). If samples from other organisms are used, a statistical analysis (e.g., average, median, geometric mean) of the methylation densities of the samples of tissue A can be used to obtain the methylation density MD_A, and similarly for MD_B.

[0063] Genomic site can be chosen to have minimal inter-individual variation, for example, less than a specific absolute amount of variation or being within a lowest portion of genomic sites tested. For instance, for the lowest portion, embodiments can select only genomic sites having the lowest 10% of variation among a group of genomic sites tested. The other organisms can be taken from healthy persons, as well as those with particular physiologic conditions (e.g. pregnant women, or people with different ages or people of a particular sex), which may correspond to a particular subpopulation that includes the current organism being tested.

[0064] The other organisms of a subpopulation may also have other pathologic conditions (e.g. patients with hepatitis or diabetes, etc.). Such a subpopulation may have altered tissue-specific methylation patterns for various tissues. The methylation pattern of the tissue under such disease condition can be used for the deconvolution analysis in addition to using the methylation pattern of the normal tissue. This deconvolution analysis may be more accurate when testing an organism from such a subpopulation with those conditions. For example, a cirrhotic liver or a fibrotic kidney may have a different methylation pattern compared with a normal liver and normal kidney, respectively. Thus, if a patient with liver cirrhosis was screened for other diseases, it can be more accurate to include a cirrhotic liver as one of the candidates contributing DNA to the plasma DNA, together with the healthy tissues of other tissue types.

B. Multiple Genomic Sites

[0065] More genomic sites (e.g., 10 or more) may be used to determine the constitution of the DNA mixture when there are more potential candidate tissues. The accuracy of the estimation of the proportional composition of the DNA mixture is dependent on a number of factors including the number of genomic sites, the specificity of the genomic sites (also called "sites") to the specific tissues, and the variability of the sites across different candidate tissues and across different individuals used to determine the reference tissue-specific levels. The specificity of a site to a tissue refers to the difference in the methylation density of the genomic sites between the particular tissue and other tissue types.

[0066] The larger the difference between their methylation densities, the more specific the site to the particular tissue would be. For example, if a site is completely methylated in the liver (methylation density = 100%) and is completely unmethylated in all other tissues (methylation density = 0%), this site would be highly specific for the liver. Whereas, the variability of a site across different tissues can be reflected by, for example, but not limited to, the range or standard deviation of methylation densities of the site in different types of tissue. A larger range or higher standard deviation would allow a more precise and accurate determination of the relative contributions of the different organs to the DNA mixture mathematically. The effects of these factors on the accuracy of estimating the proportional contribution of the candidate tissues to the DNA mixture are illustrated in the later sections of this application.

[0067] Here, we use mathematical equations to illustrate the deduction of the proportional contribution of different organs to the DNA mixture. The mathematical relationship between the methylation densities of the different sites in the DNA mixture and the methylation densities of the corresponding sites in different tissues can be expressed as:

$$\overline{MD_i} = \sum_k (p_k \times MD_{ik}),$$

where $\overline{MD_i}$ represents the methylation density of the site i in the DNA mixture; p_k represents the proportional contribution of tissue k to the DNA mixture; MD_{ik} represents the methylation density of the site i in the tissue k. When the number of sites is the same or larger than the number of organs, the values of individual p_k can be determined. The tissue-specific methylation densities can be obtained from other individuals, and the sites can be chosen to have minimal inter-individual variation, as mentioned above.

[0068] Additional criteria can be included in the algorithm to improve the accuracy. For example, the aggregated contribution of all tissues can be constrained to be 100%, i.e.

$$\sum_k p_k = 100\%.$$

Furthermore, all the organs' contributions can be required to be non-negative:

$$p_k \geq 0, \forall k$$

[0069] Due to biological variations, the observed overall methylation pattern may not be completely identical to the methylation pattern deduced from the methylation of the tissues. In

such a circumstance, mathematical analysis would be required to determine the most likely proportional contribution of the individual tissues. In this regard, the difference between the observed methylation pattern in the DNA and the deduced methylation pattern from the tissues is denoted by W.

$$W = O - \sum_k (p_k \times M_k)$$

where O is the observed methylation pattern for the DNA mixture and M_k is the methylation pattern of the individual tissue k. p_k is the proportional contribution of tissue k to the DNA mixture. The most likely value of each p_k can be determined by minimizing W, which is the difference between the observed and deduced methylation patterns. This equation can be resolved using mathematical algorithms, for example by, but not limited to, using quadratic programming, linear/non-linear regression, expectation-maximization (EM) algorithm, maximum likelihood algorithm, maximum a posteriori estimation, and the least squares method.

C. Method of Methylation Deconvolution

[0070] As described above, a biological sample including a mixture of cell-free DNA molecules from an organism can be analyzed to determine the composition of the mixture, specifically the contributions from different tissue types. For example, the percentage contribution of the cell-free DNA molecules from the liver can be determined. These measurements of the percentage contributions in the biological sample can be used to make other measurements of the biological sample, e.g., identifications of where a tumor is located, as is described in later sections.

[0071] FIG. 1 is a flowchart illustrating a method 100 of analyzing a DNA mixture of cell-free DNA molecules to determine fractional contributions from various tissue types from methylation levels according to embodiments of the present invention. A biological sample includes a mixture of cell-free DNA molecules from M tissues types. The biological sample can be any one of various examples, e.g., as mentioned herein. The number M of tissue types is greater than two. In various embodiments, M can be 3, 7, 10, 20, or more, or any number in between. Method 100 can be performed at least partially using a computer system, as can other methods described herein.

[0072] At block 110, N genomic sites are identified for analysis. The N genomic sites can have various attributes, e.g., as described in more detail in section II, which describes type I and type II

genomic sites. As examples, the N genomic sites can include type I or type II sites only, or a combination of both. The genomic sites can be identified based on analyses of one or more other samples, e.g., based on data obtained from databases about methylation levels measured in various individuals.

[0073] Specific genomic sites can be selected to provide a desired level of accuracy. For example, loci that have at least a threshold variability can be used, as opposed to just using loci that are specific to one tissue type. A first set (e.g., 10) of the genomic sites can be selected such that each have a coefficient of variation of methylation levels of at least 0.15 across M tissue types and such that each has a difference between a maximum and a minimum methylation level for the M tissue types that exceeds 0.1 for one or more other samples. This first set of genomic sites may not have a specific methylation signature for a specific tissue type, e.g., only or predominantly methylated in the specific tissue type. Such a first set is referred to as type II sites. These genomic sites can be used in combination with genomic sites that do have a specific signature, which are referred to as type I sites.

[0074] Using the type II sites can ensure that the full space of methylation levels across the tissue types is spanned by the genomic sites, thereby providing increased accuracy over the type I sites. Just using more type I sites provides redundant basis vectors for the methylation space (i.e., more genomic sites that have the same pattern as other sites), while adding other genomic sites whose methylation levels have various values across different tissues adds new basis vectors for discriminating fractional contributions via the linear system of equations.

[0075] In some embodiments, at least 10 of the N genomic sites each have a coefficient of variation of methylation levels of at least 0.15 across the M tissue types. The at least 10 genomic sites can also each have a difference between a maximum and a minimum methylation level for the M tissue types that exceeds 0.1. These methylation properties of the genomic loci can be measured for one sample or a set of samples. The set of samples may be for a subpopulation of organisms that includes the instant organism being tested, e.g., a subpopulation having a particular trait that is shared with the instant organism. These other samples can be referred to as reference tissues, and different reference tissues may be used from different samples.

[0076] At block 120, N tissue-specific methylation levels are obtained at the N genomic sites for each of M tissue types. N is greater than or equal to M, so that the tissue-specific methylation

levels can be used in the deconvolution to determine the fractional percentages. The tissue-specific methylation levels can form a matrix A of dimensions N by M. Each column of the matrix A can correspond to a methylation pattern for a particular tissue type, where the pattern is of methylation levels at the N genomic sites.

[0077] In various embodiments, the tissue-specific methylation patterns can be retrieved from public database(s) or previous studies. In examples herein, the methylation data for neutrophils and B cells were downloaded from the Gene Expression Omnibus (Hodges et al. Mol Cell 2011;44:17-28). Methylation patterns for other tissues (hippocampus, liver, lung, pancreas, atrium, colon (including its various parts, e.g. sigmoid colon, transverse colon, ascending colon, descending colon), adrenal gland, esophagus, small intestines and CD4 T cell) were downloaded from the RoadMap Epigenomics project (Ziller et al. Nature 2013;500:477-81). The methylation patterns for the buffy coat, placenta, tumor and plasma data were from published reports (Lun et al. Clin Chem. 2013;59:1583-94; Chan et al. Proc Natl Acad Sci U S A. 2013;110:18761-8). These tissue-specific methylation patterns can be used to identify the N genomic sites to be used in the deconvolution analysis.

[0078] At block 130, the biological sample including a mixture of cell-free DNA molecules from the M tissues types is received. The biological sample may be obtained from the patient organism in a variety of ways. The manner of obtaining such samples may be non-invasive or invasive. Examples of non-invasively obtained samples include certain types of fluids (e.g. plasma or serum or urine) or stools. For instance, plasma includes cell-free DNA molecules from many organ tissues, and is thus useful for analyzing many organs via one sample.

[0079] At block 140, cell-free DNA molecules from the biological sample are analyzed to identify their locations in a reference genome corresponding to the organism. For example, the cell-free DNA molecules can be sequenced to obtain sequence reads, and the sequence reads can be mapped (aligned) to the reference genome. If the organism was a human, then the reference genome would be a reference human genome, potentially from a particular subpopulation. As another example, the cell-free DNA molecules can be analyzed with different probes (e.g., following PCR or other amplification), where each probe corresponds to a genomic location, which may cover a heterozygous and one or more CpG sites, as is described below.

[0080] A statistically significant number of cell-free DNA molecules can be analyzed so as to provide an accurate deconvolution for determining the fractional contributions from the M tissue types. In some embodiments, at least 1,000 cell-free DNA molecules are analyzed. In other embodiments, at least 10,000 or 50,000 or 100,000 or 500,000 or 1,000,000 or 5,000,000 cell-free DNA molecules or more can be analyzed. The total number of molecules to analyze can depend on M and N, and the desired precision (accuracy). In various examples, the total number of cell-free DNA analyzes can be less than 500,000, one million, two million, five million, ten millsion, 20 million, or 50 million.

[0081] At block 150, N mixture methylation levels are measured at the N genomic sites using cell-free DNA molecules that are each located at any one of N genomic sites of the reference genome. A DNA molecule can be identified as located at a genomic site or a locus by one or more bases of the DNA molecule corresponding to one or more base positions of the genomic site or locus. Thus, the sequence of the DNA molecule would cover one or more base positions of the genomic site or locus. This information can be determined based on the locations determined in block 140. Such an identification of a DNA molecule located at a site of a locus can be used for any similar block of methods described herein.

[0082] The N mixture methylation levels refer to methylation levels in the mixture of the biological sample. As an example, if a cell-free DNA molecule from the mixture is located at one of the N genomic sites, then a methylation index for that molecule at the site can be included in an overall methylation density for that site. The N mixture methylation levels can form a methylation vector b of length N, where b corresponds to observed values from which the fractional contribution of each corresponding tissue type can be determined.

[0083] In one embodiment, the methylation levels for the genomic sites in the DNA mixture can be determined using whole genome bisulfite sequencing. In other embodiments, the methylation levels for the CpG sites can be determined using methylation microarray analysis, such as the Illumina HumanMethylation450 system, or by using methylation immunoprecipitation (e.g. using an anti-methylcytosine antibody) or treatment with a methylation-binding protein followed by microarray analysis or DNA sequencing, or by using methylation-sensitive restriction enzyme treatment followed by microarray or DNA sequencing, or by using methylation aware sequencing e.g. using a single molecule sequencing method (e.g. by a nanopore sequencing (Schreiber et al.

Proc Natl Acad Sci 2013; 110: 18910-18915) or by the Pacific Biosciences single molecule real time analysis (Flusberg et al. Nat Methods 2010; 7: 461-465)). Tissue-specific methylation levels can be measured in a same way. In yet other embodiments, other methods, for example, but not limited to targeted bisulfite sequencing, methylation-specific PCR, non-bisulfite based methylation-aware sequencing (e.g. by single molecule sequencing platforms (Powers et al. Efficient and accurate whole genome assembly and methylome profiling of E. coli. BMC Genomics. 2013;14:675)) can be used for the analysis of the methylation level of the plasma DNA for plasma DNA methylation deconvolution analysis.

[0084] At block 160, M values of a composition vector are determined. Each M value corresponds to a fractional contribution of a particular tissue type of the M tissue types to the DNA mixture. The M values of the composition vector can be solved to provide the N mixture methylation levels (e.g., methylation vector b) given the matrix A composed of NxM tissue-specific methylation levels (i.e. N tissue-specific methylation levels for each of the M tissue types). The M fractional contributions can correspond to a vector x that is determined by solving Ax=b. When N is greater than M, the solution can involve a minimization of errors, e.g., using least-squares.

[0085] At block 170, the composition vector is used determine an amount of each of the M tissue types in the mixture. The M values of the composition vector may be taken directly as the fractional contributions of the M tissue types. In some implementations, the M values can be converted to percentages. Error terms can be used to shift the M values to higher or lower values.

D. Applications

[0086] As mentioned above, the fractional contributions can be used in further measurements of the biological sample and other determinations, e.g., whether a particular chromosomal region has a sequence imbalance, whether a particular tissue type is diseased, and to determine which haplotype of two parental haplotypes is inherited by the fetus of a pregnant female from which the sample was obtained.

[0087] FIG. 2 shows a schematic diagram showing several potential applications of DNA methylation deconvolution (e.g., using plasma) according to embodiments of the present invention. In FIG. 2, a biological sample 205 is subjected to genome-wide bisulfite sequencing at 210. At

230, plasma DNA tissue mapping uses tissue-specific methylation profiles 220 to determine tissue contribution percentages. Example tissue-specific methylation profiles are shown as liver, blood cells, adipose tissues, lungs, small intestines, and colon. The contribution percentages can be determined as described above and elsewhere, e.g., solving Ax=b. Examples of applications include prenatal testing 241, cancer detection and monitoring 242, organ transplant monitoring, and organ damage assessment 244.

[0088] A list of methylation markers (genomic sites) that are useful for determining the contributions of different organs to the plasma DNA can be identified by comparing the methylation profiles (FIG. 2) of different tissues, including the liver, lungs, esophagus, heart, pancreas, sigmoid colon, small intestines, adipose tissues, adrenal glands, colon, T cells, B cells, neutrophils, brain and placenta. In various examples, whole genome bisulfite sequencing data for the liver, lungs, esophagus, heart, pancreas, colon, small intestines, adipose tissues, adrenal glands, brain and T cells were retrieved from the Human Epigenome Atlas from the Baylor College of Medicine (www.genboree.org/epigenomeatlas/index.rhtml). The bisulfite sequencing data for B cells and neutrophils were from the publication by Hodges et al. (Hodges et al; Directional DNA methylation changes and complex intermediate states accompany lineage specificity in the adult hematopoietic compartment. Mol Cell 2011; 44: 17-28). The bisulfite sequencing data for the placenta were from Lun et al (Lun et al. Clin Chem 2013; 59:1583-94). In other embodiments, markers can be identified from datasets generated using microarray analyses, e.g. using the Illumina Infinium HumanMethylation450 BeadChip Array.

II. SELECTION OF METHYLATION MARKERS

[0089] Above, we have described the principle of using methylation analysis to determine the composition of a DNA mixture. In particular, the percentage contribution of different organs (or tissues) to the plasma DNA can be determined using methylation analysis. In this section, we further describe the method for the selection of methylation markers and clinical applications of this technology.

[0090] The results of determining the composition of the DNA mixture by methylation analysis are affected by the methylation markers used for the deconvolution of the composition of the DNA

mixture. Thus, the selection of appropriate genomic methylation markers can be important for the accurate determination of the constitution of the DNA mixture.

A. Criteria for a Methylation Marker for Deconvolution

[0091] For marker selection, the following three attributes may be considered. (i) It is desirable for a methylation marker to have a low variability in the methylation level measured in the same tissue type across different individuals. As the determination of the composition of the DNA mixture is dependent on the recognition of the tissue-specific methylation patterns, the low variability in methylation level in the same tissue type across different individuals would be useful for accurate identification of the tissue-specific patterns in the DNA mixture. In embodiments where the tissue-specific methylation levels are obtained from samples of other organisms (e.g., from a database), the low variability means that the methylation levels from the other samples are similar to the tissue-specific methylation levels for the current organism being tested.

[0092] (ii) It is desirable for a methylation marker to have a high variability in methylation levels across different tissues. For a particular marker, a higher difference in the methylation levels across different tissues can provide a more precise determination of the contribution of different tissues to the DNA mixture. In particular, an improvement in precision can be obtained by using one set of markers having attribute (ii) and another set of markers having attribute (iii).

[0093] (iii) It is desirable for a methylation marker to have a particularly different methylation level in a particular tissue when compared with those from most or all of the other tissues. In contrast to point (ii) above, a marker can have low variability in the methylation level of most tissues but its methylation level in one particular tissue is different from most of the other tissues. This marker would be particularly useful for the determination of the contribution of the tissue that has a different methylation level from other tissues.

B. Example

[0094] A principle of marker selection is illustrated in the following hypothetical examples in table 1.

	Marker 1	Marker 2	Marker 3	Marker 4	Marker 5	Marker 6
Liver 1	20%	69%	9%	9%	10%	90%

Liver 2	50%	70%	10%	10%	10%	90%
Liver 3	90%	71%	11%	11%	10%	90%
Heart	20%	20%	30%	13%	12%	12%
Lung	30%	30%	60%	17%	14%	84%
Colon	40%	40%	90%	20%	80%	80%

Table 1. Methylation densities in different tissues for 6 hypothetical methylation markers.

[0095] In this hypothetical example, marker 2 has lower variability in methylation density in the liver from three individuals when compared with marker 1. Therefore, marker 2 is superior to marker 1 as a signature for determining the contribution of the liver in a DNA mixture.

[0096] Compared with marker 4, marker 3 has a higher variability in methylation density across different tissue types. The same level of change in the estimated contribution from the different tissues would provide a bigger change in the deduced methylation density of the DNA mixture for marker 3 than for marker 4 according to the mathematical relationship discussed above. Therefore, the estimation of the contribution of each tissue can be more precise with marker 3.

[0097] Marker 5 has a low variability in methylation density across the liver, heart and lung. Their methylation densities vary from 10% to 14%. However, the methylation density of colon is 80%. This marker would be particularly useful for determining the contribution of the colon in the DNA mixture. Similarly, the heart is hypomethylated compared with the other tissues for marker 6. Therefore, the contribution of the heart can be accurately determined by marker 6. Thus, the combination of markers 5 and 6 would be able to accurately determine the contributions of the colon and the heart. The addition of markers 2 and 3 would then be sufficient to deduce the contribution of each of the four organs, including the liver, heart, lung and colon.

C. Different Types of Markers

[0098] A methylation marker may not necessarily need to have all of the above three attributes. A type I methylation marker would typically have attribute (iii) above. A number of such markers may also have attribute (i). On the other hand, a type II methylation marker would typically have attribute (ii) above. A number of such markers may also have attribute (i). It is also possible that a particular marker may have all three attributes.

[0099] In some embodiments, markers are broadly divided into two types (type I and type II). Type I markers have tissue specificity. The methylation level of these markers for a particular group of one or more tissues is different from most of the other tissues. For example, a particular tissue can have a significant methylation level compared with the methylation level of all the other tissues. In another example, two tissues (e.g., tissue A and tissue B) have similar methylation levels, but the methylation levels of tissues A and B are significantly different from those of the remaining tissues.

[0100] Type II markers have a high inter-tissue methylation variability. The methylation levels of these markers are highly variable across different tissues. A single marker in this category may not be sufficient to determine the contribution of a particular tissue to the DNA mixture. However, a combination of type II markers, or in combination with one or more type I markers can be used collectively to deduce the contribution of individual tissues. Under the above definition, a particular marker can be a type I marker only, a type II marker only, or be simultaneously both a type I and type II marker.

1. Type I Markers

[0101] In one embodiment, a type I marker can be identified by comparing the methylation density of the marker with the mean and standard deviation (SD) of methylation densities of this particular marker for all candidate tissues. In one implementation, a marker is identified if its methylation density in one tissue is different from the mean of all the tissues by 3 standard deviations (SD).

[0102] The methylation profiles of 14 tissues obtained from sources mentioned above were studied to select markers. In one analysis, a total of 1,013 type I markers were identified (markers labeled type I in Table S1 of Appendix A of U.S. Provisional Application No. 62/158,466) using the above criteria. In other embodiments, other cutoffs between the particular tissues and the mean methylation densities can be used, for example, but not limited to 1.5 SD, 2 SD, 2.5 SD, 3.5 SD and 4 SD. In yet another embodiment, a type I marker can identified through the comparison of the methylation density of the particular tissue to the median methylation density of all tissues.

[0103] In other embodiments, the type I markers can be obtained when more than one tissue (for example, but not limited to two, three, four or five tissues) show significantly different

methylation densities than the mean methylation density of all the candidate tissues. In one implementation, a cutoff methylation density can be calculated from the mean and SD of the methylation densities of all the candidate tissues. For illustration purpose, the cutoff can be defined as 3 SD higher or lower than the mean methylation densities. A marker is selected when the methylation densities of more than one (for example, but not limited to two, three, four, five, or more than five) tissues are more than 3 SD higher than the mean methylation density or more than 3 SD lower than the mean methylation density of the tissues.

2. Type II Markers

[0104] For identification of type II markers, the mean and SD of methylation densities across all 14 candidate tissues were calculated and the ratio of SD to the mean was denoted as the coefficient of variation (CV). In this illustrative example, we used a cutoff of >0.25 for the CV to identify the qualified type II markers, as well as the difference between the maximum and minimum methylation densities for the group of tissues exceeding 0.2. Using these criteria, 5820 type II markers were identified (markers labeled type II in Table S1 of Appendix A). In other embodiments, other cutoffs for the CV, for example but not limited to 0.15, 0.2, 0.3 and 0.4, can be used. In yet other embodiments, other cutoffs for the difference between the maximum and minimum methylation densities, for example, but not limited to 0.1, 0.15, 0.25, 0.3, 0.35, 0.4, 0.45 and 0.5, can be used.

[0105] In other embodiments, the average values across multiple samples of the same tissue type can be used to measure a variation of the methylation levels across different tissues. For example, 10 methylation levels of a same genomic site from 10 samples can be averaged to obtain a single methylation level for the genomic site. A similar process can be performed to determine average methylation levels for other tissue types for the genomic site. The average values across tissue types can then be used for determining whether the genomic site has significant variation across tissue types. Other statistical values can be used besides an average, e.g., a median or a geometric mean. Such statistical values can be used to identify type I and/or type II markers.

[0106] The different samples of a same tissue type (e.g., from different individuals) can be used to determine a variation of methylation levels across the different samples. Thus, if there are multiple samples of the same tissue type, embodiments can further measure the variation of a

particular marker amongst such samples of the same tissue type. A marker with a low variation across samples would be a more reliable marker than one with a high variation. Further details of markers and deconvolution can be found in commonly-owned U.S. Patent Publication 2016/0017419, entitled "Methylation Pattern Analysis Of Tissues In A DNA Mixture," by Chiu et al., and PCT Publication WO2014/043763 entitled "Non-Invasive Determination Of Methylome Of Fetus Or Tumor From Plasma."

D. Different Categories of Markers

[0107] A "category" for a genomic locus (methylation marker) corresponds to specific variation in methylation levels for a locus across different individuals for a same tissue type. Different categories can have different ranges of variation among a particular tissue type across individuals. A first category of methylation markers might have a difference of 10% in the methylation levels or lower among the individuals tested. A second category of methylation markers might have a difference of more than 10% in the methylation levels among the individuals tested. The use of methylation markers with low inter-individual variations (first category markers) would potentially improve the accuracy of determining the contribution of the particular organ in the DNA mixture.

E. Identification of Potential Methylation Markers

[0108] In some embodiments, potential methylation markers were identified in the following manner. Such potential methylation markers can then be subjected to the above criteria to identify type I and type II markers. In other embodiments, an identification of type I or type II is not needed. And, other embodiments may use other techniques to identify potential methylation markers.

[0109] In some embodiments, all CpG islands (CGIs) and CpG shores on autosomes were considered for potential methylation markers. CGIs and CpG shores on sex chromosomes were not used so as to minimize variation in methylation levels related to the sex-associated chromosome dosage difference in the source data. CGIs were downloaded from the University of California, Santa Cruz (UCSC) database (genome.ucsc.edu/, 27,048 CpG islands for the human genome) (Kent et al., The human genome browser at UCSC, Genome Res. 2002;12(6):996-1006) and CpG shores were defined as 2 kb flanking windows of the CpG islands (Irizarry et al. The

human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores. Nat Genet 2009; 41(2):178 - 186). Then, the CpG islands and shores were subdivided into non-overlapping 500 bp units and each unit was considered as a potential methylation marker.

[0110] The methylation densities (i.e., the percentage of CpGs being methylated within a 500 bp unit) of all the potential loci were compared between the 14 tissue types. As previously reported (Lun et al. Clin Chem. 2013; 59: 1583-94), the placenta was found to be globally hypomethylated when compared with the remaining tissues. Thus, the methylation profile of the placenta was not included at the marker identification phase. Using the methylation profiles of the remaining 13 tissue types, the two types of methylation markers were identified. For example, type I markers can refer to any genomic sites with methylation densities that are 3 SD below or above in one tissue when compared with the mean level of the 13 tissue types. Type II markers can considered highly variable when (A) the methylation density of the most hypermethylated tissue is at least 20% higher than that of the most hypomethylated one; and (B) the SD of the methylation densities across the 13 tissue types when divided by the mean methylation density (i.e. the coefficient of variation) of the group is at least 0.25. Lastly, in order to reduce the number of potentially redundant markers, only one marker may be selected in one contiguous block of two CpG shores flanking one CpG island.

F. Selection based on Application

[0111] The set of methylation markers chosen for particular applications can be varied depending on the parameters of the desired applications. For example, for applications focusing on haplotype or allele analysis, useful markers would be those located on the same cell-free DNA molecules as one of the heterozygous alleles. As cell-free DNA molecules (e.g., plasma DNA) are usually less than 200 bp, useful markers can be CpG sites within 200 bp of a heterozygous locus (e.g., a SNP). As another example, for applications in which the release of DNA from a particular tissue into plasma is of special significance, one can select a preferentially larger number of methylation markers that are differentially methylated in this tissue type (e.g. type I marker) when compared with the others in the marker set.

[0112] The number and choice of methylation markers in the deconvolution analysis can be varied according to the intended use. If the fractional contribution of the liver is of particular interest, e.g. in a patient who has received a liver transplant, more type I liver specific markers can be used in the deconvolution analysis to increase the precision of the quantification of the contribution of the transplanted liver to the plasma DNA.

III. COMPOSITION ACCURACY

[0113] As described above, embodiments can identify the tissue contributors of plasma DNA. In various examples, genome-wide bisulfite sequencing of plasma DNA was performed and analyzed with reference to methylation profiles of different tissues. Using quadratic programming as an example, the plasma DNA sequencing data were deconvoluted into proportional contributions from different tissues. Embodiments were tested for pregnant women, patients with hepatocellular, lung and colorectal carcinoma, and subjects following bone marrow and liver transplantation.

[0114] In most subjects, white blood cells were the predominant contributors to the circulating DNA pool. The placental contributions in pregnant women correlated with the proportional contributions as revealed by fetal-specific genetic markers. The graft-derived contributions to the plasma in the transplant recipients correlated with those determined using donor-specific genetic markers. Patients with hepatocellular, lung or colorectal cancer showed elevated plasma DNA contributions from the organ with the tumor. The liver contributions in hepatocellular carcinoma patients also correlated with measurements made using tumor-associated copy number aberrations.

[0115] In cancer patients and in pregnant women exhibiting copy number aberrations in plasma, methylation deconvolution pinpointed the tissue type responsible for the aberrations. In a pregnant woman diagnosed as having follicular lymphoma during pregnancy, methylation deconvolution indicated a grossly elevated contribution from B-cells into the plasma DNA pool and localized B-cells (instead of the placenta) as the origin of the copy number aberrations observed in plasma. Accordingly, embodiments may serve as a powerful tool for assessing a wide range of physiological and pathological conditions based on the identification of perturbed proportional contributions of different tissues into plasma.

A. Contribution of Different Types of Blood Cells

[0116] As an example of the methylation deconvolution, we determined the contribution of different tissues and cell types to the circulating DNA. Two blood samples were collected from two patients suffering from systemic lupus erythematosus (SLE). After collection, the venous blood samples were centrifuged at 1,500 g for 10 minutes. After centrifugation, the blood cells and the plasma were separated. DNA was then extracted from the blood cells. The DNA was bisulfite converted and sequenced using one lane of a flow cell in a HiSeq2000 sequencer. Two blood cell samples were analyzed using the cell-type-specific methylation pattern analysis. The methylation patterns of neutrophils, lymphocytes, the esophagus, colon, pancreas, liver, lung, heart, adrenal glands and hippocampus were included as potential candidates of the blood cell DNA. 609 methylation markers were selected for the analysis. The whole blood samples of the two subjects were also sent for cell counting to determine the fractional composition of the neutrophils and lymphocytes of the blood cells.

	Blood sample 1		Blood sample 2	
	Cell type-specific methylation pattern analysis	Blood cell counting	Cell type-specific methylation pattern analysis	Blood cell counting
Neutrophils	90.5%	93.6%	89.4%	89.9%
Lymphocytes	9.5%	6.4%	10.6%	10.1%
Esophagus	0%	-	0%	-
Colon	0%	-	2%	-
Pancreas	0%	-	0%	-
Liver	0%	-	1%	-
Lung	1%	-	1%	-
Heart	0%	-	3%	-
Adrenal gland	0%	-	0%	-
Hippocampus	0%	-	0%	-

Table 2. Blood tissue contributions by deconvolution pattern analysis and cell counting

[0117] For methylation pattern analysis, neutrophils and lymphocytes were determined as the major components constituting the blood cell DNA. The relative proportion of the contribution of neutrophils and lymphocytes resemble their relative abundance in the blood samples according to the cell counting analysis.

B. Pregnant Women

[0118] The contributions of different tissues, including the liver, lung, pancreas, colon, hippocampus, small intestines, blood cells, heart, adrenal gland, esophagus and placenta, were analyzed using methylation analysis of the plasma DNA of pregnant women. As the placental genotype is in general identical to the fetus's genotype but different from the pregnant woman's genotype, the precise contribution of the placenta to the maternal plasma can be accurately determined by counting the number of fetal specific-alleles in the sample.

1. Composition and Correlation to Fetal DNA Percentage

[0119] Genome-wide bisulfite sequencing of plasma DNA was performed for 15 pregnant women, five from each of first, second and third trimesters. Methylation deconvolution was performed and the percentage contributions from different tissues were deduced. The contributions of different organs were determined based on the methylation levels (such as methylation densities) of all the type I and type II markers in table S1 using quadratic programming analysis.

[0120] FIG. 3A shows a graph 300 of percentage contributions of different organs to plasma DNA for 15 pregnant women according to embodiments of the present invention. Each bar corresponds to the results of one sample. The different colors represent the contributions of different organs into plasma. These results show that the white blood cells (i.e. neutrophils and lymphocytes) are the most important contributors to the plasma DNA pool. This observation is consistent with those previously obtained following bone marrow transplantation (Lui YY et al. Clin Chem 2002; 48: 421-7).

[0121] FIG. 4 shows a table 400 of percentage contributions determined from a plasma DNA tissue mapping analysis among pregnant women according to embodiments of the present invention. These results also show that the placenta is another key contributor of the plasma DNA in pregnant women, with fractional concentrations from 9.9% to 38.4%.

[0122] We also measured the placental contributions using paternally-inherited fetal single nucleotide polymorphism (SNP) alleles that were not possessed by the pregnant women as previously described (31). To analyze the fetal-specific SNP alleles, the genotypes of the fetuses were determined by analyzing the chorionic villus samples or the placenta. The genotypes of the

pregnant women were determined by analyzing the blood cells. The SNP-based results show the independent validation of the methylation deconvolution results.

[0123] FIG. 3B shows a plot 350 of a correlation between the plasma DNA fractions contributed by the placenta deduced from plasma DNA methylation deconvolution and the fetal DNA fractions deduced using fetal-specific SNP alleles according to embodiments of the present invention. Plot 350 shows that the placental contributions determined by methylation deconvolution have a strong correlation with the fetal DNA fractions measured using SNPs ($r = 0.99$, $p < 0.001$, Pearson correlation). Accordingly, a good positive correlation is observed between the values of the two parameters, suggesting that the plasma DNA methylation deconvolution accurately determines the contribution of the placenta to the maternal plasma samples.

[0124] FIG. 5 shows plots of percentage contributions of organs other than the placenta by plasma DNA tissue mapping and fetal DNA fractions based on fetal-specific SNP alleles according to embodiments of the present invention. The X-axis represents the fetal DNA fractions estimated by SNP-based analysis and the Y-axis represents the percentage contribution deduced by plasma tissue DNA mapping analysis. Plasma DNA contributions of the neutrophils show a reverse correlation. This is likely due to the fact that neutrophils are a major contributor to the plasma DNA pool and hence, as the placental contribution increases, the relative contribution from the neutrophils would by necessity decrease. The methylation deconvolution results of the remaining tissues show no correlation with the fetal DNA fraction.

[0125] FIG. 6 shows a table 600 of percentage contributions from plasma DNA tissue mapping analysis among the non-pregnant healthy control subjects according to embodiments of the present invention. When the process was applied to plasma of non-pregnant healthy controls, placental contribution was absent in most samples (median: 0%; interquartile range: 0% to 0.3%).

2. Comparison of Selected Markers vs. Random Markers

[0126] The accuracy of the percentage contributions was tested with select markers relative to random markers. Different composition calculations were done for different sets of markers. One set was chosen based on criteria mention above, and the other was a random set. The results show

that it is important to judicially choose the methylation markers (genomic loci) use, in order to obtain accurate results.

[0127] Eleven pregnant women and four healthy non-pregnant subjects were recruited for this analysis. Their plasma DNA was bisulfite converted and sequenced using the Illumina HiSeq2000 sequencer. Each plasma sample was sequenced with one lane of a sequencing flow cell. The sequence reads were then analyzed using a bioinformatic program, Methy-Pipe (Jiang P. PLoS One 2014; 9: e100360). This program can align the bisulfite-converted sequence reads to the reference genome and determine the methylation status of each CpG site on each sequenced fragment.

[0128] The first set of markers have high specificity for identifying the different tissues in the plasma DNA. For each tissue type, markers that have the biggest difference in methylation density compared with the other tissues were selected. The markers were determined from genomic regions containing at least one CpG dinucleotide. In this example, CpG islands (CGIs) were used as potential markers, having a high frequency of CpG sites in a particular stretch of DNA. CGIs in this particular example are downloaded from the University of California, Santa Cruz (UCSC) database: (genome.ucsc.edu). In total, we obtained 27,048 CpG islands from the human genome. The median size of a CpG island is 565 bp (range: 200 bp to 45 kb). 90% of the islands are less than 1.5 kb.

[0129] For each methylation marker, the difference in methylation density between the tissue-of-interest and the other tissues was determined. The difference is then expressed as the number of standard deviations (SDs) across the other tissues. For the tissue-of-interest, all the markers were ranked according to this difference in methylation density. The 20 markers with the biggest difference above (10 markers) and below (10 markers) the mean methylation densities of the other tissues were selected. The number of markers can vary, for example, but not limited to 5, 15, 20, 30, 40, 50, 100 and 200.

[0130] In addition, markers with a high variability across all the different tissues were also selected. In this example, markers with >50% difference between the tissues with the highest and lowest methylation densities were selected. In other applications, other values, for example, but not limited to 20%, 30%, 40%, 60%, 70% and 80%, can be used. Furthermore, the variability of methylation densities across different tissues was also calculated based on the mean and SD. In

this example, a marker was also selected if the value of SD is more than two times the mean. In other applications, other cutoff values, for example, but not limited to 1, 1.5, 2.5 and 3, can also be used. Based on these selection criteria, 344 methylation markers were selected for the first set.

[0131] For the second set, 341 markers were randomly selected from the 27,048 CGIs discussed above. All the CGIs were first numbered from 1 to 27,048. Then a random number (between 1 and 27,048) was generated by a computer for marker selection. This process was then repeated until a total of 341 markers were selected. If a random number generated had been used, another one would be generated. This set of markers is expected to have a much lower specificity in identifying the tissue-specific methylation patterns. Thus, the accuracy of determining the composition of the plasma DNA is expected to be reduced.

[0132] FIG. 7 shows a table 700 of the estimated contributions of different organs to the plasma DNA for 11 pregnant women and 4 non-pregnant healthy subjects using the first set of markers (with high organ specificity) according to embodiments of the present invention. The fetal DNA fractions were determined by counting fetal-specific alleles and are shown in the bottom row. In each of the four non-pregnant control subjects, the contribution of the placenta to the plasma was determined to be close to 0%. This indicates the specificity of this approach.

[0133] FIG. 8 shows a table 800 of the estimated contributions of different organs to the plasma DNA for 11 pregnant women and 4 non-pregnant healthy subjects using the second set of markers (with low organ specificity) according to embodiments of the present invention. The fetal DNA fractions determined by counting fetal-specific alleles are shown in the bottom row. Using these less specific markers, a relatively non-concordant percentage of contribution from the placenta was observed, and considerable contributions from the placenta were observed in the four non-pregnant control subjects. This indicates that the tissue specificity of the markers is important in this approach.

[0134] FIG. 9A is a plot 900 showing the correlation between the estimated fetal DNA fraction (contribution from the placenta) and the fetal DNA fraction determined by counting the fetal-specific alleles in the maternal plasma samples. The results from the two techniques have good correlation using the first set of methylation markers. However, using the second set of methylation markers, the estimation by using the methylation analysis showed significant deviation from the true values determined using fetal-specific alleles counting.

[0135] FIG. 9B is a plot 950 showing absolute difference between the estimation from methylation markers and fetal DNA fraction determined by fetal-specific alleles counting. The median error of the estimation using methylation analysis were 4% and 8% using the first set of markers and the second set of markers, respectively.

C. Effect of Different Criteria

[0136] As described above, various criteria can be used to identify markers of different types. For example, a type I marker can be identified by a methylation level in a particular tissue that is different from the mean methylation level for all tissues, e.g., at least by a specific threshold, such as 3 SD. And, for type II markers, criteria of a certain variation and maximum difference are used. Sections below show accuracy of different criteria for identifying markers.

1. Performance of Markers with Less Stringent Criteria

[0137] We compared the performance of methylation deconvolution analysis using markers with different variability across different tissues. The placental contributions to plasma DNA were determined for 15 pregnant women based on two sets of markers with different selection criteria. Both sets of markers include all the type I markers as described in previous sections. However, the selection criteria of type II markers are different for the two sets of markers.

[0138] Set I markers include all the 5,820 type II markers fulfilling the criteria of having methylation density CV >0.25 and the difference between the maximum and minimum methylation densities for the groups of tissues exceeding 0.2. For Set II markers, the CV requirement was >0.15 and the difference between the maximum and minimum methylation densities for the groups of tissues exceeded 0.1. There were 8,511 type II markers in this set of markers.

[0139] FIG. 10A is a graph 1000 showing placental contribution to plasma DNA deduced using markers with different selection criteria according to embodiments of the present invention. The vertical axis corresponds to placental contribution deduced using the set II markers. The horizontal axis corresponds to placental contribution deduced using the set I markers. There was a good correlation between the placental contribution results based on the two sets of markers with different selection criteria (r=0.99, Pearson correlation). Accordingly, good accuracy can be

obtained using the requirements of CV >0.15 and of the difference between the maximum and minimum methylation densities for the groups of tissues exceeding 0.1.

2. Effect of Methylation Level Variation within Same Type of Tissue

[0140] To investigate if the variation in methylation level of markers between the same type of tissues (e.g. from different individuals) would affect the performance of deconvolution analysis, we analyzed placental tissues from two pregnant cases. Two categories of methylation markers were identified. Specifically, the two categories were identified based on their similarity in methylation levels in two placental tissues. Markers of category i have a methylation density of 10% or lower. Markers of category ii have high variability between the two placental tissues (difference in methylation density of more than 10%).

[0141] FIG. 10B is a graph 1050 showing the accuracy of plasma DNA deconvolution using markers with low variability (category i) and high variability (category ii) in the same type of tissue. Plasma DNA deconvolution was performed to determine the placental contribution to the plasma DNA for 15 pregnant women. For each marker, the mean of the methylation densities of the two placental tissues were used to represent the methylation level of the placenta in the analysis. For each of the deconvolution analysis using the category i and category ii markers, a total of 1,024 markers were used.

[0142] The amount of placentally-derived DNA in plasma was further determined based on the proportion of the fetal-specific SNP alleles. The percentage contribution deduced by the methylation deconvolution analysis based on category i and category ii markers were then compared with the results based on fetal-specific SNP alleles. The median deviation of the derived placental contribution from the value estimated based on fetal-specific alleles was 2.7% and 7.1% using category i and category ii markers, respectively. Thus, the use of category i markers which had lower inter-individual variation in the tissue methylation level gave better accuracy in the methylation deconvolution analysis.

[0143] Significantly higher difference between the results from methylation deconvolution and fetal-specific allele analysis was observed when markers with high variability within the same type of tissue (category ii) were used ($P<0.0001$, Wilcoxon sign-rank test). In other words, the use of markers with low variability within the same type of tissue would increase the accuracy of

methylation deconvolution analysis. Accordingly, markers can be selected based on the variability within the same type of tissues, for example, but not limited to the value of CV and the difference between the maximum and minimum methylation density for the same type of tissues.

IV. DECONVOLUTION OF FETAL SIGNATURES

[0144] If a genomic signature (e.g., a particular SNP allele) is known, embodiments can determine which tissue is the origin of such signatures. Thus, if a particular signature is representative of a fetus (e.g., a paternal allele at a particular locus), then the fractional contribution for the signature would be substantial for the placental tissue.

[0145] To illustrate that single nucleotide alteration can also be used to determine the tissue of origin that the alteration is derived from, we analyzed the plasma DNA of a pregnant woman. The placenta and the maternal buffy coat were genotyped to identify the SNPs that the mother was homozygous and the fetus was heterozygous. We denote the allele shared by the fetus and the mother as A and the fetal-specific allele as B. Therefore, the mother had a genotype of AA and the fetus had a genotype of AB at each of these SNPs.

[0146] After bisulfite sequencing of the maternal plasma DNA, all DNA fragments carrying the fetal-specific allele (B allele) and at least one CpG site were selected and used for downstream analysis. A total of 1.31 billion fragments were sequenced and 677,140 fragments carrying the fetal-specific allele (B allele) were used for the deconvolution analysis. All CpG sites that were covered by at least 10 DNA fragments were used for deconvolution analysis. Other numbers of DNA fragments covering a site can be used, such as 5, 15, 20, 25, or 30. As the B allele was fetal specific, these DNA fragments were expected to be derived from the placenta.

Tissue	Contribution (%)
Liver	0.9
Lung	0.0
Colon	0.0
Small intestines	0.0
Pancreas	0.5
Adrenal glands	0.0
Esophagus	3.1
Adipose tissues	0.0
Heart	0.0
Brain	0.3

T cells	0.0
B cells	0.0
Neutrophil	0.0
Placenta	95.2

Table 3. Methylation deconvolution analysis using fetal-specific allele.

[0147] In Table 3, from the methylation deconvolution analysis, it was shown that the placenta was deduced to be the major contributor for these DNA fragments carrying fetal-specific SNP alleles. These results suggest that the methylation deconvolution analysis accurately identified the tissue origin of these DNA fragments carrying fetal-specific alleles.

[0148] This shows that a particular allele can be attributed to a fetus. Such a technique is described in more detail below for determining genotypes and haplotypes of a fetus using methylation deconvolution analysis.

V. DETERMINATION OF FETAL GENOME (MUTATIONAL ANALYSIS)

[0149] For noninvasive prenatal testing, the analysis of the inheritance of a maternal mutation using maternal plasma DNA is a challenging task. For example, if a pregnant woman is heterozygous for a mutation, the analysis on the mutational status of the fetus using maternal plasma DNA analysis would be technically difficult because both the mutant and the normal alleles would be present in her plasma, regardless of the mutational status of her fetus. Previously, a number of different approaches have been developed to address this problem (Lun et al. Proc Natl Acad Sci U S A. 2008;105:19920-5; Lo et al. Sci Transl Med. 2010;2:61ra91; Lam et al. Clin Chem. 2012;58:1467-75). The principle of these previous approaches involves the comparison between the relative amounts of the mutation and the normal allele in maternal plasma. To enhance the statistical power of the comparisons, some of these approaches further involve the comparisons of the relative amounts of SNP alleles linked to the mutation and those linked to the normal allele. As an alternative or in addition, some embodiments of the present invention can deduce the mutational status of the fetus by methylation deconvolution analysis.

A. Contribution for Alleles using Methylation Deconvolution

[0150] In this example, a genotype of the fetus is determined. Assume the genotypes of the father and mother are NN and MN at a particular locus, respectively. M and N denote the mutant

and normal alleles, respectively. In this scenario, the fetus can inherit either the M allele or the N allele from the mother. Therefore, there are two possible genotypes for the fetus, namely MN and NN. In maternal plasma, DNA carrying the fetal genotype is actually derived from the placenta. Thus, these DNA fragments would exhibit the placental methylation profile.

[0151] FIG. 11A shows a first scenario where the fetus has inherited the M allele from the mother and has a genotype of MN at a particular locus according to embodiments of the present invention. In the top part of FIG. 11A (labeled Genotypes), the father is shown as having genotype NN, the mother is shown as having genotype MN, and the fetus is shown as having genotype MN. The DNA fragments that exhibit the placental methylation profile are marked with a P, where are shown on the fetal genotype. For example, the placental methylation profile can correspond to certain methylation levels at genomic sites near the particular locus. DNA fragments that align to the particular locus can also include genomic sites near the locus (e.g., within 200 bp of the locus), and thus can be used to measure methylation levels for the methylation deconvolution analysis. Considering the genotypes of the parents, the M allele is specific for the mother and the N allele is shared between the father and the mother.

[0152] In the bottom part of FIG. 11A (labeled Maternal plasma), instances of the two alleles M and N are shown, with each instance representing a different DNA molecule in the plasma at the locus of interest. Only a small number of DNA molecules are shown for illustration purposes. In this example, the fetal DNA percentage is assumed to be 25%, as shown by 25% of the DNA molecules being marked with a P.

[0153] In the maternal plasma sample, we selectively analyzed the DNA fragments carrying the M allele and performed the methylation deconvolution analysis. Because the fetus has a genotype of MN, the placenta would contribute both M and N alleles to the maternal plasma DNA. Therefore, some of the DNA fragments carrying the M allele would also carry the placenta-specific methylation profile at genomic sites near the locus. The methylation deconvolution analysis would indicate that some of the DNA fragments carrying the M allele would be derived from the placenta, and thus the fetal genotype does include the M allele.

[0154] FIG. 11B shows a second scenario where the fetus has inherited the N allele from the mother and has the genotype of NN at a particular locus according to embodiments of the present invention. In this situation, only the DNA fragments carrying the N allele would exhibit the

placental methylation profile in maternal plasma. Thus, the selective analysis of the DNA fragments carrying the M allele with methylation deconvolution would indicate that these DNA fragments do not have significant contribution from the placenta. Accordingly, it can be determined that the fetus does not have M, and thus has a genotype of NN.

[0155] In some embodiments, the placental contribution for the M and N alleles can be compared. Here, we assume that the fetal DNA accounts for approximately 10% of the total maternal plasma DNA. The selective deconvolution of the M and N alleles would be useful to indicate which allele the fetus has inherited from the mother. The expected results are shown in Table 4 below:

	Fetal genotype	
	MN	NN
Placental contribution for plasma DNA carrying the M allele	Approximately 10%	Insignificant (close to zero)
Placental contribution for plasma DNA carrying the N allele	Approximately 10%	Approximately 20%
Ratio of placental contribution for M and N alleles (M:N)	1:1	0:2

Table 4. Placental contributions for M and N alleles for NN paternal genotype.

[0156] In Table 4, the percentage placental contribution of the M and N alleles can be compared. An approximately equal placental contribution for the two alleles (e.g., within a threshold of each other) suggests that the fetal genotype is MN. On the other hand, a significantly higher placental contributions for the N allele compared with the M allele would indicate a fetal genotype of NN.

[0157] In another embodiment, the paternal genotype does not need to be taken into account. In this situation, the possible genotypes of the fetus include MM, MN, and NN.

	Fetal genotype		
	MN	NN	MM
Placental contribution for plasma DNA carrying the M allele	Approximately 10%	Insignificant (close to zero)	Approximately 20%
Placental contribution for plasma DNA carrying the N allele	Approximately 10%	Approximately 20%	Insignificant (close to zero)

Ratio of placental contribution for M and N alleles (M:N)	1:1	0:2	2:0

Table 5. Placental contributions for M and N alleles for unknown paternal genotype.

[0158] In table 5, the placental contribution for the DNA fragments carrying the M and N alleles for different fetal genotypes are shown. When the fetus has a genotype of MM, the placental contribution for the M allele would be significantly higher than that for the N allele. When the fetus has a genotype of NN, the placental contribution for the N allele would be significantly higher than that for the N allele. When the fetus has a genotype of NM, the placental contribution for the M allele would be approximately equal to the placental contribution for the N allele.

[0159] Accordingly, where the paternal genotype is not known, fractional contributions can be determined for both alleles. That is, a first fractional contribution can be determined using a first set of cell-free DNA molecules that align to the locus and include N. Methylation levels of the first set of cell-free DNA molecules can be measured at K genomic sites near the locus. And, a second fractional contribution can be determined using a second set of cell-free DNA molecules that align to the locus and include M. Methylation levels of the second set of cell-free DNA molecules can be measured at the K genomic sites near the locus. For the first scenario of the fetal genotype being MN, the fractional contributions determined for either allele would be about the same, as can be tested to by determining whether the fractional contributions are within a threshold value of each other.

[0160] To illustrate the feasibility of this approach, we analyzed the plasma DNA of a pregnant woman. The plasma DNA was bisulfite converted and analyzed using massively parallel sequencing. In addition, the placenta and blood cells were analyzed to determine the genotype of the fetus and the mother. For illustration purposes, a SNP located within the *KLF2* gene was analyzed. For this SNP, the genotypes of the mother and the fetus were CG and CC, respectively. With this genotype combination, the placenta would contribute the C allele to the maternal plasma, but all the G alleles in the maternal plasma would be derived from the maternal tissues.

[0161] In the sequencing data, there were 24 fragments carrying the G allele and 55 fragments carrying the C allele. The CpG sites within these DNA fragments were used for methylation deconvolution. In this analysis, an objective is to determine the placental contribution of the two

alleles. To illustrate the principle, only the placenta and the blood cells were considered as candidate tissues for the methylation deconvolution analysis. In another embodiment, three or more types of tissues can be used as candidates. In yet another embodiment, tissues expected to have significant contribution, for example blood cells, liver, lung, intestines and placenta, can be used as candidates.

	C allele	G allele	C/G ratio
Placenta	62.6%	1.8%	34
Blood cells	37.4%	98.2%	

Table 6. Placental contributions for C and G alleles for unknown paternal genotype.

[0162] In Table 6, the contribution from the placenta was deduced to be 62.6% and 1.8% for the C allele and the G allele, respectively. The ratio of placental contribution for C/G is 34. These results suggest that the genotype of the fetus would be CC. This is consistent with the genotyping result of the placental tissue.

[0163] This embodiment is different from and potentially has more utility than a previous method for noninvasive prenatal testing based on the analysis of allelic ratio for DNA with a specific methylation pattern (Tong et al. Clin Chem 2006; 52: 2194-202). In this previous method, tissue-specific DNA is first identified from a DNA mixture (e.g. plasma DNA) based on methylation pattern. For example, a particular gene is completely unmethylated in the blood cells and methylated in the placenta. The identification is performed using an enzyme that leaves the methylated placental DNA intact.

[0164] Thus, all the methylated DNA molecules remaining in the plasma would be derived from the placenta rather than from the blood cells. Then, the allelic ratio for a SNP located on the placenta-derived DNA molecules can be determined by measuring amounts of the different alleles at the locus using the intact placental DNA. When the fetus is heterozygous for the SNP, the ratio of the two alleles in the placenta-specific DNA would be approximately 1. However, if the fetus is affected by an aneuploid chromosome and has three copies of the chromosome carrying this particular SNP, the ratio of the two alleles would be either 1:2 or 2:1.

[0165] In this previous method, the tissue-specific DNA molecules need to be first identified based on a methylation status that is unique to the tissue of interest. The methylated DNA molecules are unique for the placenta because the blood cells are completely unmethylated for the

targeted region. However, in this present embodiment, the uniqueness of a certain methylation state is not required. The candidate tissues only need to be different in their methylation profiles, accordingly more loci can be used, thereby enabling haplotype deconvolution. Thus, the tissue contributions can be determined for the different alleles based on their methylation profiles. Further, the previous method may be more susceptible to statistical variations as the numbers of fetal reads with each allele are compared directly to each other. Whereas, when the placental contributions are compared to each other, the numbers of fetal reads are not compared directly to each other. Instead, the placental contribution is determined from all of the reads (methylated or not), and thus the placental contributions can be the same, even when the number of fetal reads differs. Thus, a coverage bias to one haplotype can be accounted for.

B. Determination of Inherited Haplotype using Deconvolution

[0166] It has previously been demonstrated that through the analysis of plasma DNA (or other cell-free DNA) of a pregnant woman carrying a fetus, the maternal haplotypes inherited by the fetus can be deduced using the process of relative haplotype dosage analysis (RHDO) (Lo et al. Sci Transl Med 2010; 2: 61ra91 and U.S. Patent 8,467,976). In this method, one uses the haplotype information for the pregnant woman. This latter information can be obtained using family analysis or a method for the direct analysis of the haplotype (e.g. Fan et al. Nat Biotechnol 2011; 29: 51-57; Snyder et al. Nat Rev Genet 2015; 16: 344-358). SNPs that are heterozygous in the mother but homozygous in the father can be used for the RHDO analysis. Such a use of specific SNPs can limit the loci that can be used, and therefore limit the amount of data and accuracy. Embodiments may not be so restricted to such specific SNPs. Further, embodiments can be used in combination with the above references to provide increased accuracy.

[0167] Embodiments can use methylation deconvolution to determine placental contributions using the cell-free DNA molecules for two haplotypes. The placental contributions can be compared to determine which haplotype is inherited by the fetus. Embodiments can start with deduced maternal or paternal haplotypes, and then measure the methylation levels of plasma DNA molecules containing SNP alleles in each of those deduced haplotypes. One can then perform methylation deconvolution. The fetal haplotype can be identified as the one with the highest placental contribution from the methylation deconvolution analysis. In all of the above embodiments, the paternal or maternal haplotypes can, instead of being a deduced one, also be

determined by family analysis (i.e. by analyzing the DNA of other family members) or by a direct method (e.g. the method described by Fan et al Nat Biotechnol 2012).

1. Maternal Haplotypes

[0168] In this example, we demonstrate plasma DNA methylation deconvolution analysis can be used for deducing the maternal haplotypes inherited by an unborn fetus. A source of genomic DNA from the pregnant woman, e.g. the buffy coat DNA, can be subjected to genotyping, e.g. using a microarray. Then, the maternal genotyping results are entered into a haplotype deduction program (e.g., IMPUTE2, Howie et al. PLoS Genet. 2009;7:e1000529) to deduce the likely first maternal haplotype and the second maternal haplotype. Population-specific genotype and haplotype information can be taken into consideration for improving the accuracy of deduction. In other embodiments, the parental haplotypes can be worked out by single molecule analysis, for example but not limited to the methods described by Fan et al (Nat Biotechnol. 2011;29:51-7) , Kaper et al (Proc Natl Acad Sci U S A. 2013;110:5552-7), Lan et al. (Nat Commun. 2016;7:11784) and Selvaraj et al (Nat Biotech 2013;31:1111–1118). Then, maternal plasma DNA can be subjected to genome-wide bisulfite sequencing and alignment to reference genomic sequences. Methylation deconvolution can then be performed for each of the predicted haplotypes. As fetal DNA in maternal plasma is predominantly of placental origin, the maternal haplotype inherited by the fetus is the one that shows the highest placental contribution.

[0169] The maternal haplotype information can be used to link the SNP alleles and the CpG sites on the same homologous chromosome together. Then, DNA fragments from the same chromosome copy (haplotype) can be identified using the SNP alleles. The CpG sites (or other sites) on this particular chromosome copy (haplotype) can be used for the methylation deconvolution. As the number of CpG sites that can be used for deconvolution would be proportional to the number of SNPs on the homologous chromosome and much bigger than the number of CpG sites linked to a single SNP in the haplotype-based deconvolution analysis, this method would be more precise than the deconvolution analysis using CpG site(s) that are linked to a single SNP. The principle is illustrated in FIG. 12A.

[0170] FIG. 12A shows a determination of a maternal haplotype inherited by a fetus using methylation deconvolution according to embodiments of the present invention. In the top part of

FIG. 12A, the two haplotypes of the mother and the fetus are shown at three loci that the mother is heterozygous. The two maternal haplotypes are labeled as Hap I and Hap II. In this example, the fetus has inherited Hap I from the mother. For illustration purpose, only the SNP loci that the mother is heterozygous are shown. For illustration purposes, the father is homozygous for each of these loci in this example. However, the same principle extends to scenarios that the father is heterozygous without any change.

[0171] In the bottom part of FIG. 12A (labeled Maternal plasma), instances of the two alleles at each locus are shown, with each instance representing a different DNA molecule in the plasma at the locus of interest. Only a small number of DNA molecules are shown for illustration purposes. In this example, the fetal DNA percentage is assumed to be 20%, as shown by 20% of the DNA molecules being marked with a P.

[0172] In maternal plasma, the DNA molecules carrying the fetal genotype are derived from the placenta and therefore carrying the placental specific methylation patterns. The circles labeled with "P" represent the CpG sites exhibiting the placental methylation pattern near the heterozygous locus. A read including a heterozygous locus and a neighboring site can be used for measuring a methylation level for detecting the placental methylation pattern. In this example, an objective is to determine if the fetus has inherited Hap I or Hap II from the mother. To achieve this, plasma DNA fragments that carry alleles on Hap I and cover at least one CpG site are selected for methylation deconvolution. As the fetus has inherited Hap I from the mother, the placenta would contribute a significant proportion to this pool of plasma DNA molecules. On the other hand, when the fragments carrying alleles on Hap II are analyzed with methylation deconvolution, a very low contribution from the placenta would be observed.

[0173] To illustrate this, we analyzed the maternal plasma sample stated above for Table 6. We focused on a 5-Mb region on chromosome 1. SNPs where the mother was heterozygous and the fetus was homozygous were selected for the analysis. For each of these SNP loci, the alleles that were shared between the mother and the fetus formed one haplotype (denoted as Hap I) and the alleles that were present only on the maternal genome formed another haplotype (denoted as Hap II). Thus, in this example, there are two maternal haplotypes (Hap I and Hap II) and the fetus has inherited Hap I from the mother. In the maternal plasma, the DNA fragments carrying the alleles on Hap I and those carrying the alleles on Hap II were analyzed separately using methylation

deconvolution. All the CpG sites on the same plasma DNA molecule of a heterozygous SNP were used for the deconvolution analysis. In this example, none of these CpG sites overlapped with type I or type II markers.

	Hap I	Hap II
Liver	0%	0%
Lung	0%	6.7%
Colon	3.4%	6.2%
Small intestines	0%	10.6%
Pancreas	4.1%	0%
Adrenal glands	0%	4.6%
Esophagus	0%	0%
Adipose tissues	3.7%	3.6%
Heart	0%	0%
Brain	6.8%	10.6%
T cells	6.8%	21%
B cells	8.9%	11.7%
Neutrophil	12.7%	25%
Placenta	53.5%	0%

Table 7. Methylation deconvolution for Hap I and Hap II.

[0174] Table 7 shows deconvolution of plasma DNA fragments carrying the alleles on the two maternal haplotypes, namely Hap I and Hap II. The fetus had inherited the maternal Hap I. From this deconvolution analysis, the placenta was deduced to contribute 53.5% of the plasma DNA fragments carrying the alleles on Hap I. On the other hand, there was no contribution from the placenta to the plasma DNA fragments carrying the alleles on Hap II. Therefore, the methylation deconvolution analysis had accurately predicted that the fetus had inherited Hap I from the mother. Greater accuracy may be achieved using CpG sites that overlap with type I and/or type II markers.

[0175] As a further example, to demonstrate the practical utility of this approach, another pregnant woman was recruited. Maternal peripheral blood was taken. The blood sample was fractionated into plasma and the cellular components. The maternal buffy coat was analyzed using an Illumina HumanOmni2.5-8 BeadChip array. We used IMPUTE2 (Howie et al. PLoS Genet. 2009;7:e1000529) to deduce the phase of 851 heterozygous SNPs on a 5 Mb region on the

telomeric end of chromosome 1p. The haplotype phasing was based on reference haplotypes of 1000 genomes (mathgen.stats.ox.ac.uk/impute/1000GP_Phase3.tgz).

[0176] After the phased haplotypes were obtained, the CpG sites linked to the two haplotypes were used to perform methylation deconvolution. All the CpG sites on the same plasma DNA molecule of a heterozygous SNP were used for the deconvolution analysis. In this example, none of these CpG sites overlapped with type I or type II markers. Among the 851 SNPs used for the deconvolution, 820 (96.2%) were on intron and intergenic regions. None of them overlapped with CpG islands or shores.

	Hap I	Hap II
Liver	0	0
Lung	0	5.4
Colon	0	6.2
Small intestine	0	0
Pancreas	0	25
Adrenal glands	0	0
Esophagus	0	0
Adipose tissues	0	17.8
Heart	0	0
Brain	0	0
T cells	11	7.9
B cells	0	0
Neutrophils	20.2	28.4
Placenta	68.9	9.3

Table 8. Methylation deconvolution for Hap I and Hap II.

[0177] Table 8 shows a deconvolution of plasma DNA fragments carrying the alleles on the two maternal haplotypes deduced from a panel of reference haplotypes. The two haplotypes are named Hap I and Hap II. The deduced Hap I has significantly higher amount of placental contribution than Hap II, namely, 68.9% versus 9.3%. Thus, the maternal Hap I was deduced to have been inherited by the fetus. The maternal inheritance relied on haplotype deduction was consistent with results from maternal and fetal genotypes.

[0178] An advantage of this method is that one is not restricted to SNPs for which the father of the fetus is homozygous and the mother of the fetus is heterozygous. Indeed, in the above example,

we had performed the analysis without knowing or deducing the paternal genotype or haplotype. This is an advantage over previously described methods ((Lo et al. Sci Transl Med 2010; 2: 61ra91, US Patent 8,467,976, Fan et al. Nature 2012; 487: 320- 324, Kitzman et al. Sci Transl Med 2012; 4: 137ra76).

[0179] In some embodiments, a first fractional contribution for a first haplotype can be compared to a reference value derived based on the fetal DNA fraction to determine whether the haplotype has been inherited by the fetus. The cutoffs can be calculated as for example but not limited to 1 time, 1.2 times, 1.4 times, 1.6 times, 1.8 times, 2 times, 2.2 times, 2.4 times, 2.6 times or 2.8 times of the fetal DNA fraction. In this manner, the second fractional contribution for a second haplotype does not need to be determined, if the first fractional contribution is sufficiently large.

[0180] In some embodiments, the inherited haplotype may have a deconvoluted fractional concentration double that of the fetal fraction and the non-inherited one has an insignificant contribution. The contribution of the non-inherited haplotype may not have a zero contribution as the paternal haplotypes may give noise to this analysis because some paternal alleles may be the same as the maternal alleles. If the level of noise is high, the fractional contribution of the second haplotype can be determined, and the one with a higher deconvoluted fraction can be deduced to be inherited by the fetus.

[0181] Some implementations could test both haplotypes using the reference value, to confirm that only one is inherited. If both appear to be inherited, then the two fractional contributions can be compared to each other. Additionally, if both appear to be inherited, the paternal genome can be checked, as the fetus could have inherited a paternal haplotype that matches the non-inherited maternal haplotype.

[0182] In other embodiments, the second fractional contribution can be used to determine the reference value, e.g., the second fractional contribution plus a threshold value. Thus, the reference value can be a sum of the second fractional contribution and a threshold value.

2. Paternal haplotype

[0183] In another embodiment, methylation deconvolution analysis can be applied for the analysis of the paternal haplotype inheritance.

[0184] FIG. 12B shows an illustration of the paternal haplotype methylation analysis according to embodiments of the present invention. Methylation deconvolution can be performed on the maternal plasma DNA fragments carrying the alleles on paternal Hap III and Hap IV. As Hap III has been inherited by the fetus, the placental contribution would be higher for Hap III compared with Hap IV. Thus, the paternal inheritance of the fetus can be deduced.

[0185] This embodiment has advantages over previous methods based on the analysis of paternal-specific alleles. For example, for SNP at position 1, the A allele is present in the father, but not in the mother. Therefore, the detection of the paternal-specific A allele in maternal plasma indicates the inheritance of the Hap III by the fetus. However, for SNP at position 2, both the C and T alleles are not fetal-specific. In this situation, paternal-specific allele analysis cannot be used. However, the methylation deconvolution analysis does not require the presence of paternal-specific allele. Thus, SNPs that are heterozygous in both the father and the mother can be used for the methylation deconvolution analysis of the two paternal haplotypes.

[0186] Accordingly, a similar process, as used for maternal haplotypes, can be used to determine which paternal haplotype is inherited. In FIG. 12B, the placental contribution for Hap III would be higher than the placental contribution from Hap IV. The paternal haplotypes can be determined in a same or similar manner as the maternal haplotypes can be determined.

3. Method using Deconvolution

[0187] FIG. 13 is a flowchart illustrating a method 1300 of determining a portion of a fetal genome from a maternal sample using methylation deconvolution according to embodiments of the present invention. The biological sample includes a mixture of cell-free DNA molecules from a plurality of tissues types, including maternal tissue types and a fetal tissue type. The fetus has a father and a mother being the pregnant female. The portion of the fetal genome can be an entire chromosome copy or just part of the chromosome copy. The determined portions of the fetal genome can be combined to provide information on different portions of the fetal genome, up to the entire fetal genome.

[0188] At block 1310, a plurality of cell-free DNA molecules from the biological sample are analyzed. Block 1310 can be performed using techniques described in block 140 of method 100 of FIG. 1. For example, at least 1,000 cell-free DNA molecules can be analyzed to determine where

the cell-free DNA molecules are located, and methylation levels can be measured as described below. Further, the cell-free DNA molecules are analyzed to determine a respective allele of the cell-free DNA molecule. For example, an allele of a DNA molecule can be determined from a sequence read obtained from sequencing or from a particular probe that hybridizes to the DNA molecule, where both techniques can provide a sequence read (e.g., the probe can be treated as the sequence read when there is hybridization).

[0189] At block 1320, a first haplotype and a second haplotype of a first chromosomal region of a first parental genome of a first parent of the fetus are determined. One skilled in the art will be aware of various techniques to determine haplotypes of a parent. The haplotypes may be determined from the same sample as used to determine methylation levels below or from a different sample. In some implementations, the haplotypes can be determined from cellular samples, e.g., the buffy coat of a blood sample or the tissue of another organ. Examples of determining haplotypes are provided in U.S. Patent No. 8,467,976, which is incorporated by reference in its entirety. The first parent can be the mother or the father. Other examples of methods for detecting the parental haplotypes include, but not limited to the methods described by Fan et al (Nat Biotechnol 2011; 29: 51-57), Snyder et al (Nat Rev Genet 2015; 16: 344-358), the GemCode technology from 10X Genomics (www.10xgenomics.com/), and Targeted Locus Amplification (TLA) technology from Cergentis (www.cergentis.com/).

[0190] At block 1330, one or more heterozygous loci are identified from the first and second haplotypes. Each heterozygous locus has a corresponding first allele in the first haplotype and a corresponding second allele in the second haplotype. The one or more heterozygous loci may be a first plurality of heterozygous loci, where a second plurality of heterozygous loci can correspond to a different chromosomal region.

[0191] At block 1340, a first set of the plurality of cell-free DNA molecules is identified. Each of the plurality of cell-free DNA molecules is located at any one of the heterozygous loci from block 1330 and includes a corresponding first allele, so that the cell-free DNA molecule can be identified as corresponding to the first haplotype. It is possible for a cell-free DNA molecule to be located at more than one of the heterozygous loci, but typically a read would only include one heterozygous locus. Each of the first set of cell-free DNA molecules also includes at least one of N genomic sites, where the genomic sites are used to measure the methylation levels. N is an integer,

e.g., greater than or equal to 2, 3, 4, 5, 10, 20, 50, 100, 200, 500, 1,000, 2,000, or 5,000. Thus, a read of a cell-free DNA molecule can indicate coverage of 1 site, 2 sites, etc.

[0192] At block 1350, N first mixture methylation levels are measured at the N genomic sites (e.g., CpG sites) using the first set of the plurality of cell-free DNA molecules. One first mixture methylation level can be measured for each of the N genomic sites. Block 1350 can be performed in a similar manner as block 150 of method 100 of FIG. 1. In some embodiments, the measurement of the methylation level of a DNA molecule can use methylation-aware sequencing results, which may also be used to determine the location and respective allele of the DNA molecule. One skilled in the art will be aware of the various techniques that can be used to determine a methylation status of a site on a DNA molecule.

[0193] At block 1360, a first fractional contribution of the fetal tissue type in the mixture is determined using the N first methylation levels. In some embodiments, block 1360 can be performed via blocks 160 and 170 of method 100 of FIG. 1. Thus, a fractional contribution can be determined simultaneously for a panel of M tissue types. Block 1360 may use N issue-specific methylation levels at N genomic sites, determined for each of M tissue types, e.g., as in block 120 of method 100 of FIG. 1.

[0194] At block 1370, a second set of the plurality of cell-free DNA molecules is identified. Each of the plurality of cell-free DNA molecules is located at any one of the heterozygous loci from block 1330 and includes a corresponding second allele, so that the cell-free DNA molecule can be identified as corresponding to the second haplotype. Each of the second set of cell-free DNA molecules also includes at least one of the N genomic sites, where the genomic sites are used to measure the methylation levels.

[0195] At block 1380, N second mixture methylation levels at the N genomic sites are measured using the second set of the plurality of cell-free DNA molecules. Block 1380 may be performed in a similar manner as block 1350.

[0196] At block 1385, a second fractional contribution of the fetal tissue type in the mixture is determined using the N second methylation levels. Block 1385 may be performed in a similar manner as block 1360.

[0197] At block 1390, a first separation value is computed between the first fractional contribution and the second fractional contribution. Examples of separation values are described herein, e.g., including a difference or a ratio.

[0198] At block 1395, the portion of fetal genome is determined at the one or more heterozygous loci based on the first separation value. Thus, an inherited haplotype of the first parent can be determined. For example, the first separation value can be a ratio of the first fractional contribution and the second fractional contribution. The portion of the fetal genome can be determined to have one or more copies of the first haplotype and no copies of the second haplotype when the ratio is greater than a threshold value. Examples of threshold values include but not limited to 1.3, 1.4, 1.5, 1.6, 1.8, 2.0, 2.2, 2.4, 2.6, 2.8 and 3.0. The portion of the fetal genome can be determined to have one or more copies of the second haplotype and no copies of the first haplotype when the ratio is less than a threshold value. Examples of threshold values include but not limited to 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7 and 0.8. The portion of the fetal genome can be determined to have the first haplotype and the second haplotype when the ratio is equal to one within a cutoff value. Examples of cutoff values include but not limited to 0.85, 0.9, 0.95, 1.0, 1.05, 1.1 and 1.15. Both haplotypes might be inherited when both parents have a same haplotype in the region being analyzed.

[0199] As another example, the first separation value is a difference of the first fractional contribution and the second fractional contribution. The portion of the fetal genome can be determined to have one or more copies of the first haplotype and no copies of the second haplotype when the difference is greater than a threshold value. Examples of threshold values include but not limited to 1%, 1.5%, 2%, 2.5%, 3%, 4%, 5%, 6%, 7%, 8%, 10%, 12%, 14%, 16%, 18% and 20%. The portion of the fetal genome can be determined to have one or more copies of the second haplotype and no copies of the first haplotype when the difference is less than a threshold value, e.g., where the threshold value is a negative number.

[0200] The inherited haplotype of the other parent can also be determined. For example, a second plurality of heterozygous loci of the first chromosomal region can be identified in the genome of the other parent. Fractional contributions can be determined for each of the other parent's haplotypes, and a separation value can be used to determine the inherited haplotype of the other parent.

[0201] For example, the first plurality of heterozygous loci and the second plurality of heterozygous loci can be the same loci or be different. Each of the second plurality of heterozygous loci can include a corresponding third allele in a first haplotype of the other parent (e.g., a first paternal haplotype) and a corresponding fourth allele in a second haplotype of the other parent (e.g., a second paternal haplotype). The third and fourth alleles can be the same as the first and second alleles. In addition to the first and second set of cell-free DNA molecules for the first parent, a third set of the plurality of cell-free DNA molecules can each be located at any one of the second plurality of heterozygous loci, include the corresponding third allele of the heterozygous locus, and include at least one of K genomic sites. The K genomic sites may be the same or different than the N genomic sites used for the first parent. In a similar manner as with the first parent, K third mixture methylation levels can be measured at the K genomic sites using the third set of the second plurality of cell-free DNA molecules, and a third fractional contribution of the fetal tissue type in the mixture can be determined using the K third methylation levels. The third fractional contribution corresponds to the first haplotype of the other parent (e.g., the first paternal haplotype).

[0202] A fourth set of the plurality of cell-free DNA molecules can each be located at any one of the second plurality of heterozygous loci, include the corresponding fourth allele of the heterozygous locus, and include at least one of the K genomic sites. Thus, the fourth set of DNA can be used to test the second haplotype of the other parent. K fourth mixture methylation levels at the K genomic sites can be measured using the fourth set of the second plurality of cell-free DNA molecules, and a fourth fractional contribution of the fetal tissue type in the mixture can be determined using the K fourth methylation levels. A second separation value can be computed between the third fractional contribution and the fourth fractional contribution, and the portion of fetal genome at the second plurality of heterozygous loci can be determined based on the second separation value. The inherited haplotype from the other parent can be determined in a similar manner as for the first parent. The fourth fractional contribution corresponds to the second haplotype of the other parent (e.g., the second paternal haplotype).

[0203] In some embodiments, the second fractional contribution does not need to be determined. Instead, a haplotype can be determined to be inherited if the corresponding fractional contribution is sufficiently high. For example, the first fractional contribution can be compared to a reference

value to determine whether the fetus inherited the first haplotype at the first chromosomal region. The fetus can be determined to have inherited the first haplotype at the first chromosomal region when the first fractional contribution exceeds the reference value.

[0204] In other embodiments, the reference value may be determined from the second fractional contribution. For example, the reference value can be a sum of the second fractional contribution and a threshold value. The sum with the threshold value can ensure that the first fractional contribution is sufficiently greater than the second fractional contribution.

[0205] A separate determination of inheritance can be made for the second haplotype by comparing the second fractional contribution to the reference value to determine whether the fetus inherited the second haplotype at the first chromosomal region. The fetus can be determined to have inherited the second haplotype at the first chromosomal region when the second fractional contribution exceeds the reference value. If both the fractional contributions are determined to exceed the reference value, the two fractional contributions can be compared to each other to determine if one is significantly greater than the other (e.g., using a threshold). The haplotypes of the other parent can be determined to identify whether one of these haplotypes is the same as the haplotypes of the first parent, thereby explaining that both haplotypes of the first parent could have been inherited.

C. Determination of Inherited Haplotype using Methylation Levels

[0206] Other embodiments can use the general hypomethylation of cell-free fetal DNA to identify the inherited haplotype as the one with the lower overall methylation level. Embodiments can start with deduced maternal or paternal haplotypes, and then measure the methylation levels of plasma DNA molecules containing SNP alleles in each of those deduced haplotypes. In one implementation of analyzing the maternal haplotypes, the methylation levels of the two deduced maternal haplotypes can be compared, and the one with the lower methylation level would be predicted to be the haplotype inherited by the fetus. In another implementation of analyzing the paternal haplotypes, the methylation levels of the two deduced paternal haplotypes can be compared, and the one with the lower methylation level would be predicted to be the haplotype inherited by the fetus.

1. Example

[0207] As an example, a methylation level of each of the two maternal haplotypes can be determined. As the placental tissue is relatively hypomethylated compared with other tissues, it is expected that the maternal haplotype inherited by the fetus would be more hypomethylated than the one that is not inherited by the fetus. The methylation densities were tested in the maternal plasma using the actual haplotypes of the mother, which were deduced using the maternal, paternal, and fetal genotypes.

	Hap I	Hap II
Overall methylation densities	65%	87%

Table 9. Methylation densities for actual Hap I and Hap II.

[0208] Table 9 shows methylated densities of the two maternal haplotypes in the maternal plasma. As Hap I was the actual haplotype inherited by the fetus by genotyping, the results of the methylation analysis of the haplotype correctly identified the inheritance.

[0209] In other embodiments, the maternal haplotypes can be deduced based on the genotypes of the mother alone, or reference haplotypes of the population from haplotype database can also be used for this analysis. The maternal haplotypes used in this example were phased using the IMPUTE2 program. Thus, deduced maternal haplotypes can also be used in this analysis.

	Hap I	Hap II
Overall methylation densities	68%	76%

Table 10. Methylation densities for deduced Hap I and Hap II.

[0210] Table 10 shows methylated densities of the two deduced maternal haplotypes in the maternal plasma. The deduced maternal haplotype that was inherited by the fetus had lower methylation densities. An example of a statistical procedure that one can use to determine whether one haplotype has a sufficiently lower methylation density includes the chi-square test. A separation between the two methylation levels can be required to be sufficiently large (e.g., greater than a threshold) to make the determination. If the separation is not sufficient, then an indeterminate classification can be made. In some embodiments, a determination of an inheritance of both haplotypes can be determined if the separation is not sufficiently large and if both methylation levels are below a threshold level, which may be characterized by inclusion of fetal

DNA. For example, tables 9 and 10 indicate that a methylation density below 70% may indicate that the fetus has inherited that haplotype. Both haplotypes may be inherited when the parents share a haplotype for the region being analyzed.

[0211] In another embodiment, the overall methylation densities of the maternal plasma DNA carrying the paternal Hap III and Hap IV can be compared. Similar to the maternal haplotype analysis, the fetus would be deduced as having inherited the paternal haplotype that has lower overall methylation densities.

2. Method using Methylation Levels

[0212] FIG. 14 is a flowchart illustrating a method 1400 of determining a portion of a fetal genome from a maternal sample using methylation levels according to embodiments of the present invention. The biological sample includes a mixture of cell-free DNA molecules from a plurality of tissues types, including maternal tissue types and a fetal tissue type. The fetus has a father and a mother being the pregnant female. The portion of the fetal genome can be an entire chromosome copy or just part of the chromosome copy. The determined portions of the fetal genome can be combined to provide the entire fetal genome, as with other methods described herein.

[0213] At block 1410, a plurality of cell-free DNA molecules from the biological sample are analyzed. Block 1410 can be performed in a similar manner as block 1310 of method 1300 of FIG. 13.

[0214] At block 1420, a first haplotype and a second haplotype of a first chromosomal region are determined of a first parental genome of a first parent of the fetus. Block 1420 can be performed in a similar manner as block 1320 of FIG. 13. In some embodiments, genotypes of the first parent's genome can be determined at the plurality of heterozygous loci using a sample from the first parent, e.g., a blood sample or other tissue that may or may not include fetal DNA. A plurality of reference haplotypes can be obtained, e.g., from databases of reference genomes. The first haplotype and the second haplotype can be deduced using the genotypes and the plurality of reference haplotypes. For example, the alleles of each genotype can be compared against the reference haplotypes, and any haplotypes that do not include the alleles at the corresponding locus can be discarded. Once two reference haplotypes remain, those haplotypes can be identified as being the first haplotype and the second haplotype.

[0215] At block 1430, a plurality of heterozygous loci is identified from the first and second haplotypes. Each heterozygous locus has a first allele in the first haplotype and a second allele in the second haplotype.

[0216] At block 1440, a first set of the plurality of cell-free DNA molecules is identified. Block 1440 can be performed in a similar manner as block 1340 of FIG. 13.

[0217] At block 1450, a first mixture methylation level is measured using the first set of the plurality of cell-free DNA molecules. For example, the first mixture methylation level may be a methylation density for the cell-free DNA molecules of the first set. The methylation density can be computed as a total methylation density for all of the cell-free DNA molecules of the first set. In another example, separate methylation densities can be computed for each locus, and the separate methylation densities can be combined to obtain the first mixture methylation level, e.g., an average of the separate methylation densities.

[0218] At block 1460, a second set of the plurality of cell-free DNA molecules is identified. 1460 can be performed in a similar manner as block 1370 of FIG. 13.

[0219] At block 1470, a second mixture methylation level is measured using the second set of the plurality of cell-free DNA molecules. For example, the second mixture methylation level may be a methylation density for the cell-free DNA molecules of the second set.

[0220] At block 1480, it is determined which of the first haplotype and the second haplotype is inherited by the fetus based on which of the first mixture methylation level and the second mixture methylation level is lower. As part of block 1480, a separation value can be determined between the first mixture methylation level and the second mixture methylation level, and compared to a threshold value. The threshold value can ensure that the lower level is sufficiently lower. The threshold value can be determined using the chi-square test. For example, measurements can be taken of samples where the inherited haplotype is known, and a distribution of the separation values can be determined, and a threshold value can be selected that accurately determines the inherited haplotype in the training data obtained from the samples. Methods 1300 and 1400 can also be combined, with each being performed as a check, and the inherited haplotype determined if both methods are consistent with each other.

D. Selection of loci

[0221] Various embodiments can be used for the comparison of the methylation levels or the fraction contributions of the two deduced maternal haplotypes in the maternal plasma. In one embodiment, the number of SNP loci to be analyzed can be determined before the analysis. For example, the number of SNP loci used in the haplotype deconvolution analysis can be determined according to a number of factors, for example, but not limited to the desired statistical power, the mean difference in methylation levels in the placenta and blood cells in the region of interest, and the number of molecules being analyzed for each SNP.

[0222] The size of the region of interest can be fixed, and all SNPs within the region of interest can be used in the analysis. The size of the region of interest can be determined taking into account a number of factors, for example, but not limited to the desired statistical power, the mean difference in methylation level in the placenta and blood cells in the region of interest, the number of molecules being analyzed for each SNP, and the chance of meiotic recombination with the region of interest.

[0223] In other embodiments, the number of SNPs and the size of the region to be analyzed are not determined before the analysis. For example, the number of SNPs can be sequentially increased until the data are sufficient to arrive at a statistically significant conclusion concerning which maternal haplotype is statistically significantly less methylated than the other one. For instance, the SNPs on the region of interest can be arranged in an ascending order of their genomic coordinates. Then statistical testing can be carried out with the data of the SNP with the lowest number of genomic coordinate. If this is sufficient to make a conclusion regarding which haplotype is less methylated statistically, then a conclusion is made. Similarly, the SNPs can be arranged in a descending order, with a highest number of genomic coordinate that is sufficient being used.

[0224] If the statistical accuracy is not sufficient, another statistical comparison can be performed starting from the next SNP with a higher number of genomic coordinates. On the other hand, if the data of the first SNP are not sufficient for one to conclude that one haplotype is less methylated than the other one (or that the separation value between fractional contributions is not sufficiently large), the data of another SNP can be added and another round of statistical testing is

carried out. This procedure can be continued until the accumulated data are sufficient to make a statistically significant conclusion. A number of statistical tests can be performed to compare the methylation levels of the two haplotypes, for example, but not limited to Student's t-test, Mann-Whitney rank-sum test and Chi-square test. The level of statistical significance can be determined based on the desired confidence of the conclusion, for example, but not limited to adopting a P-value of 0.05, 0.01, 0.001, 0.0001, or 0.00001.

E. Combinations with RHDO

[0225] In some embodiments, results generated by the RHDO analysis of U.S. Patent 8,467,976 can be combined with present methylation embodiments to arrive at a more accurate procedure for diagnosis or to reduce an amount of sequencing required. For example, fetal haplotypes can be determined using present embodiments and using results of the RHDO analysis of U.S. Patent 8,467,976, and the determined fetal haplotypes from both of the techniques can be compared. For example, the results from the two analyses would be accepted only if they are concordant. Further analysis can be performed if the two analyses show different conclusions, e.g., measurements can be repeated at higher depth of coverage on the genome.

[0226] For such a combined approach to be most cost-effective, it is preferred to have one type of sequencing that can yield data for both methods. In one embodiment, this can be done by a single molecule method that would generate sequencing as well as methylation information, e.g. using the Single Molecule Real Time sequencing technology from Pacific Biosciences, or nanopore sequencing (e.g. from Oxford Nanopore Technologies). These are two examples of methylation-aware sequencing. In another embodiment, the RHDO analysis can be performed on bisulfite sequencing results. For such an embodiment, any maternal and paternal genetic information can be determined also using bisulfite sequencing. Bisulfite sequencing is thus another example of methylation-aware sequencing. Furthermore, other methylation aware sequencing technologies can be used, such as oxidative bisulfite sequencing (Booth et al. Science 2012; 336: 934-937) or Tet-assisted bisulfite sequencing (Yu et al. Cell 2012; 149: 1368-1380). The latter examples would allow one to analyze the 5-methylcytosine distribution of the analyzed DNA molecules.

F. Uses of knowledge of fetal genome

[0227] The noninvasive prenatal analysis of the fetal genome can be used to determine if a fetus has inherited a disease from the parents. This is particularly useful for the detection of monogenic diseases, for example congenital adrenal hyperplasia (New et al. J Clin Endocrinol Metab 2014;99:E1022-30), beta-thalassemia (Lam et al. Clin Chem. 2012;58:1467-75) and hereditary muscular dystrophies (Genet Med 2015;17:889-96). If a monogenic disease is detected, various treatments can be performed, e.g., the pregnancy can be terminated, treatment provided before pregnancy, or after birth. For example, steroid treatment can be given prenatally to a pregnant woman confirmed of having a fetus affected by congenital adrenal hyperplasia to avoid abnormal sexual development.

VI. HAPLOTYPE DECONVOLUTION ANALYSIS FOR ANEUPLOIDY DETECTION

[0228] The haplotype deconvolution can also be used to detect a sequence imbalance of a chromosomal region of a fetus, such as aneuploidies, microdeletions, or microamplifications (e.g., microduplications). For example, a fractional contribution of a haplotype in one region can be compared to a fractional contribution of another haplotype in another region.

A. Mother

[0229] FIG. 15 shows a chromosomal aneuploidy detection based on haplotype deconvolution for maternal haplotypes according to embodiments of the present invention. In this illustration, the mother has two maternal haplotypes, namely Hap I and Hap II. For illustration purposes, we assume that 80% of her plasma DNA was derived from her own cells and 20% was derived from the placenta, which are example percentages in commonly measured ranges. This method can be generally applied for pregnancies with different fetal DNA percentages. Knowledge of the fetal DNA percentage is not required, but is simply provided for illustration, although a measurement of the fetal DNA percentage may be performed in various ways, e.g., using fetal-specific alleles or fetal-specific methylation markers.

[0230] The fetus has inherited Hap I and another haplotype from the father, namely Hap III. The placentally-derived DNA would exhibit the fetal genotypes, and a sequence imbalance can be detected by analyzing the fractional contribution resulting from the placentally-derived DNA.

[0231] As illustrated above, the fetal inheritance of the maternal haplotype can be determined through the deconvolution of the two maternal haplotypes. Analysis for the placental contribution to the maternal DNA can be performed for each of the two maternal haplotypes. The maternal haplotype inherited by the fetus (Hap I in this example) would have a much higher placental contribution compared with the maternal haplotype that is not inherited by the fetus (Hap II). The placental contribution for Hap I would be positively correlated with the fetal DNA fraction in the maternal plasma.

[0232] After determining which maternal haplotype has been inherited by the fetus, the dosage of the chromosome the fetus has inherited from the mother can be further determined through maternal haplotype deconvolution. In this illustration, two chromosomal regions are analyzed using maternal haplotype deconvolution. In one embodiment, the reference chromosome (RefChr) is a chromosome or a chromosomal region which is unlikely to be affected by a chromosomal aneuploidy. The reference chromosomal region is shown on the left side of FIG. 15. The target chromosome (TargetChr) is a chromosome or a chromosomal region which is potentially affected by a chromosomal aneuploidy. The target chromosomal region is shown on the right side of FIG. 15. The two regions can be for different regions of a same chromosome or for regions of two different chromosomes.

[0233] In the example shown, the fetus has been deduced to have inherited Hap I from the mother for both the reference chromosome and target chromosome through the methylation deconvolution of the Hap I and the Hap II at each region. Then, the placental contribution to maternal plasma DNA for Hap I can be compared between the reference chromosome and the target chromosome. If the placental contribution of Hap I for the target chromosomal region is significantly different than the placental contribution of Hap I for the reference chromosomal region (e.g., higher for amplification or lower for deletion), then a sequence imbalance can be identified.

[0234] For illustration purpose, we use the detection of trisomy as an example. However, other types of chromosome aneuploidies, including monosomy, amplification of a subchromosomal region or deletion of a subchromosomal region can also be detected using this method. For trisomy, the extra copy of the affected chromosome can be inherited from the father (denoted as Trisomy (F)) or the mother (denoted as Trisomy (M)). In over 90% of trisomy 21 cases, the extra

copy of chromosome 21 is derived from the mother (Driscoll et al. N Engl J Med 2009;360: 2556–2562). In the scenario of Trisomy (M), the placental contribution of Hap I for the target chromosome would be higher than that for the reference chromosome. In FIG. 15, Trisomy (M) is shown with two instances of Hap I, which would provide a higher placental contribution for the target region than the one instance of Hap I for the reference region.

[0235] Whether the placental contribution of Hap I for the target chromosome is higher than that for the reference chromosome can be determined by comparing a separation value between the two placental contributions and a threshold, which may be based on a separate measurement of the fetal DNA percentage. A higher fetal DNA percentage would result in a higher expected separation value between the two placental contributions, and thus the threshold can be set higher. For example, with the fetal DNA percentage being 20%, the placental contribution of Hap I for the reference region would be about 20% and the placental contribution of Hap I for the target region would be about 36.4%.

[0236] For instance, assume that 10 DNA molecules exist at the reference chromosome, then two of them are fetal and eight of them are maternal. For the two fetal DNA molecules, one is derived from Hap I and one is derived from Hap III. For the eight maternal DNA molecules, four are Hap I and four are Hap II. For the target region, there would be an extra DNA molecule of Hap I from the fetus. Thus there would be two fetal Hap I DNA molecules and 4 maternal Hap I DNA molecules total, providing 2/6=33.3%. The threshold value for the difference (e.g., 13.3%) can be placed between 0 and 13.3% to provide optimal specificity and sensitivity. A distribution of separation values can be determined from a reference group of samples. In the scenario of Euploid, the placental contributions would be approximately equal, e.g., the separation value would be less than the threshold. One skilled in the art will know how to select a suitable threshold based on the description herein and in U.S. Patent 8,467,976, and other references cited herein.

[0237] In one embodiment, the ratio (or other separation value) of placental contribution of Hap I between the target and reference chromosomes for a group of pregnant women, each known to be carrying a euploid fetus can be used as a reference interval. The ratio in the tested case can be compared with this reference group to determine if a significant elevation of placental contribution of Hap I is present for the target region relative to the reference region. In the example of 20% fetal DNA, the ratio would be 33.3/20 = 1.67. The ratio can be generalized to 2/(1+f), where f

represents the fetal DNA fraction. In another embodiment, the difference in the placental contribution of Hap I between the target and the reference chromosomes can be determined. This difference is then compared with a reference group.

B. Father

[0238] In another embodiment, haplotype deconvolution of the paternal haplotypes (Hap III and Hap IV) can be performed in the maternal plasma. The analysis of the paternal haplotypes can be performed in a similar manner as for the maternal haplotypes.

[0239] FIG. 16 shows a chromosomal aneuploidy detection based on haplotype deconvolution for paternal haplotypes according to embodiments of the present invention. In this illustration, the father has two paternal haplotypes, namely Hap III and Hap IV. As in FIG. 15, the fetus has inherited Hap I from the mother and Hap III from the father.

[0240] In the scenario where the extra copy of chromosome is derived from the father (Trisomy (F)), the placental contribution of Hap III would be higher for the target chromosome than for the reference chromosome. This is shown for the Trisomy (F) example, where two copies of Hap III are shown. As described above for the maternal haplotypes, a separation value between the placental contributions of Hap III for the target and reference regions can be compared to a threshold to determine whether an extra copy of Hap III exists for the target region. In various embodiments, a ratio or difference of the two placental contributions of the tested case can be compared with a reference group of pregnant women, each known to be carrying a euploid fetus so as to determine if the fetus has a chromosomal trisomy for the target chromosome, or amplification or deletion of a target chromosomal region. The threshold can be based on the separation values for the reference group of euploid fetuses, a reference group of aneuploidy fetuses, or both. A separate measure of fetal DNA percentage can also be used, as is described herein.

C. Method of detecting a sequence imbalance

[0241] FIG. 17 is a flowchart of a method 1700 for detecting a sequence imbalance in a portion of a fetal genome of an unborn fetus of a pregnant female using a biological sample from the pregnant female according to embodiments of the present invention.

[0242] At block 1710, a plurality of cell-free DNA molecules from the biological sample are analyzed. Block 1710 can be performed in a similar manner as block 1310 of method 1300 of FIG. 13.

[0243] At block 1720, a first target haplotype of a target chromosomal region of a first parental genome of a first parent of the fetus is determined, and a first reference haplotype of a reference chromosomal region of the first parental genome is determined. Block 1720 can be performed in a similar manner as block 1320 of FIG. 13. The target chromosomal region and the reference chromosomal region can be an entire chromosome or just part of a chromosome. Thus, the target chromosomal region can be a first chromosome and the reference chromosomal region can be a second chromosome different from the first chromosome. The first parent can be the mother or the father of the fetus.

[0244] The target chromosomal region can be selected based on various criteria. For example, a plurality of target regions can be selected, as may occur to test many non-overlapping regions of a specified size, such as 1 Mb, 5 Mb, 10 Mb, 20 Mb, 50 Mb, etc. As another example, the target chromosomal region can be selected based on a copy number analysis that identifies the region as having more DNA molecules than expected, e.g., as described in U.S. Patent Publications 2009/0029377 and 2011/0276277.

[0245] In some embodiments, it can be determined that the fetus has inherited the first target haplotype from the first parent and that the fetus has inherited the first reference haplotype from the first parent. The determination can include embodiments of FIG. 13 or FIG. 14. For example, determining that the fetus has inherited the first target haplotype from the first parent can include determining a second target fractional contribution of the fetal tissue type in the mixture corresponding to the second target haplotype, computing a second separation value between the first target fractional contribution and the second target fractional contribution, and determining that the fetus has inherited the first target haplotype from the first parent based on the second separation value.

[0246] At block 1730, a plurality of target heterozygous loci are identified of the target chromosomal region of the first parental genome. Each target heterozygous locus includes a corresponding first target allele in the first target haplotype and a corresponding second target allele in a second target haplotype of the first chromosomal region of the first parental genome.

Referring back to the example of FIG. 15, the target heterozygous loci have corresponding first target alleles of {G,T,A} on Hap I and have corresponding second target alleles of {A,G,C} on Hap II.

[0247] At block 1740, a target set of the plurality of cell-free DNA molecules is identified. Each cell-free DNA molecule of the target set is located at any one of the target heterozygous loci, includes a corresponding first target allele, and includes at least one of N genomic sites in the target chromosomal region. Block 1740 can be performed in a similar manner as described herein. For example, sequence reads can be mapped to a reference genome, where the target set of plurality of cell-free DNA molecules aligns to any one of the target heterozygous loci.

[0248] At block 1750, N first mixture methylation levels are measured at the N genomic sites using the target set of the plurality of cell-free DNA molecules. Block 1750 can be performed in a similar manner as block 1350 of FIG. 13.

[0249] At block 1760, a first fractional contribution of the fetal tissue type in the mixture is determined using the N first methylation levels. Block 1760 can be performed in a similar manner as block 1360 of FIG. 13.

[0250] At block 1770, a plurality of reference heterozygous loci are identified for the reference chromosomal region of the first parental genome. Each reference heterozygous locus includes a corresponding first reference allele in the first reference haplotype and a corresponding second reference allele in a second reference haplotype of the reference chromosomal region of the first parental genome. Referring back to the example of FIG. 15, the reference heterozygous loci have corresponding first target alleles of {A,T,C} on Hap I and have corresponding second target alleles of {T,C,A} on Hap II.

[0251] At block 1775, a reference set of the plurality of cell-free DNA molecules is identified. Each cell-free DNA molecule of the reference set is located at any one of the reference heterozygous loci, includes a corresponding first reference allele, and includes at least one of K genomic sites in the reference chromosomal region.

[0252] At block 1780, K reference mixture methylation levels are measured at the K genomic sites using the reference set of the plurality of cell-free DNA molecules.

[0253] At block 1785, a first reference fractional contribution of the fetal tissue type is determined in the mixture using the K reference methylation levels.

[0254] At block 1790, a first separation value is computed between the first target fractional contribution and the first reference fractional contribution.

[0255] At block 1795, the first separation value is compared to a threshold value to determine a classification of whether the fetus has a sequence imbalance for the target chromosomal region. If the first separation value exceeds the threshold value then a sequence imbalance can be identified. The threshold value can be determined as described above, e.g., based on separation values seen in a reference group of samples not having a sequence imbalance and/or a reference group of samples having the sequence imbalance. As examples, the classification can be positive, negative, or indeterminate for the sequence imbalance being tested.

[0256] Different threshold values can be used, depending on the type of sequence imbalance. For example, if the sequence imbalance is a deletion, then the first separation value would be expected to be a negative value. In such a case, the threshold value can be a negative number, and the comparison can determine that the first threshold value exceeds the threshold value by being a larger negative number. If the sequence imbalance being tested is an amplification, then it can be tested whether the separation value is greater than the threshold value. Thus, the threshold value used can be dependent on the type of sequence imbalance being tested.

VII. DECONVOLUTION OF SIGNATURES TO IDENTIFY DISEASED TISSUE

[0257] If a genomic signature (e.g., a particular SNP allele) is known, embodiments can determine which tissue is the origin of such signatures. As the cell-free DNA molecules exhibiting the signatures are from the tissue of origin, the tissue of origin can be identified from the fractional contributions determined using cell-free DNA molecules exhibiting the signatures. Thus, cell-free DNA molecules with a signature of a transplanted organ (e.g., a signature of a haplotype of the transplanted organ) can be used to monitor changes in amounts of cell-free DNA molecules from the transplanted organ with high sensitivity, e.g., given that a high fractional contribution of the DNA in the mixture would be from the transplanted organ. Examples are provided for transplants to show that the technique is accurate. In another example, a signature of a tumor can be used to identify tissues within which the tumor resides.

A. Organ Transplantation

[0258] As an example for organ transplantation, we analyzed the plasma of a patient who had received liver transplantation and a patient who had received bone marrow transplantation. For each case, the donor-specific SNP alleles were identified through the genotyping of the tissues from the patients and the donors. For the liver transplant recipient, a biopsy of the donor liver and the blood cells of the recipient were sequenced. For the bone marrow transplant case, the buccal swab (recipient genotype) and the blood cells (donor genotype) were sequenced. The plasma DNA samples were sequenced after bisulfite conversion. Sequenced DNA fragments carrying a donor-specific SNP allele and at least one CpG site were used for downstream methylation deconvolution analysis. A total of 72 million and 121 million reads were sequenced for the patients who had received liver and bone marrow transplantation, respectively. For the two cases, 38 and 5355 fragments were used for deconvolution analysis, respectively.

Tissue type	Liver transplant recipient	Bone marrow transplant recipient
Liver	45.4	4.4
Lung	0.0	1.5
Colon	29.3	6.3
Small intestines	0.0	1.8
Pancreas	0.0	0.0
Adrenal glands	0.0	0.0
Esophagus	0.0	0.0
Adipose tissues	0.0	14.8
Heart	0.0	0.0
Brain	14.5	9.6
T-cells	0.0	12.3
B-cells	5.9	16.6
Neutrophils	4.9	32.8

Table 6. Fractional contributions for different organs to plasma DNA fragments carrying donor-specific alleles in the two transplant recipients.

[0259] Table 6 shows methylation deconvolution analysis on plasma DNA fragments carrying donor-specific alleles in a liver transplant recipient and a bone marrow transplant recipient. The

numbers represent the percentage contribution of different tissues to the donor-specific plasma DNA fragments. For the liver transplant case, the liver was shown to be the most important contributor to these DNA fragments. For the bone marrow transplant case, the hematopoietic system (including the T-cells, B-cells and neutrophils) was the major contribution of the donor-specific DNA fragments. These results indicate that methylation deconvolution can accurately indicate the tissue origin of DNA fragments having single nucleotide alterations. A small amount of sequenced fragments were attributed to other tissues probably because of measurement imprecision as relatively small number of donor-specific fragments were used for the deconvolution analysis.

[0260] The fractional contributions for tissue associated with the transplanted organ can be determined in the above manner and monitored. With the baseline fractional contribution (an example of a reference fractional contribution) relatively high as a result of using only cell-free DNA molecules exhibiting the donor signature, small changes in total amount of donor DNA in plasma can be detected. Accordingly, methylation deconvolution analysis can be applied for the monitoring of organ transplantation.

[0261] As can be seen above for the liver transplant, methylation deconvolution is not absolutely specific. In this analysis, plasma DNA fragments carrying the donor-specific alleles were used for methylation deconvolution analysis. These fragments are specific for the donor and should be derived only from the liver in this liver transplant recipient. Therefore, the theoretical contribution of the liver should be 100%. Another possibility is that certain cell types are present in different types of tissues making the liver methylation profile overlapping with other tissues. For example, the connective tissue cells in the liver may also be present in other organs. But, relative percentages from other patients or other samples (e.g., at other times) of the instant patient can identify whether more cell-free DNA molecules are being released.

[0262] In various embodiments, the donor signature can correspond to a particular haplotype of the donor genome or both haplotypes in a chromosomal region. Methylation deconvolution can be performed using cell-free DNA molecules located on the particular donor haplotypes, and increases in the fractional contribution of the particular haplotype can be monitored. If a significant increase occurs (e.g., as measured by a percentage or absolute threshold), then a rejection of the transplanted organ can be identified.

[0263] FIG. 18 shows an illustration of haplotype deconvolution for organ transplant monitoring according to embodiments of the present invention. The donor has haplotypes labeled Hap I and Hap II, and the recipient has haplotypes labeled Hap III and Hap IV. The donor has a signature at locus 1 and locus 3, as the alleles are not found on the recipient haplotypes. Locus 2 and locus 4 do not have a donor signature. Thus, embodiments may use DNA molecules that are located at locus 1 and locus 3 as part of a deconvolution process.

[0264] The plasma DNA deconvolution can be used to determine whether the determined fractional contribution from the transplanted organ is at a baseline or increased relative to the baseline. In some embodiments, the fractional contributions can be determined for each of Hap I and Hap II separately, if different signatures exist; such different signatures can exist at different loci. In other embodiments, a single fractional contribution can be determined for both haplotypes, e.g., when they share a signature. In the example shown in FIG. 18, Hap I and Hap II do share a signature at locus 1 and locus 3.

[0265] Accordingly, the contribution of the transplanted organ can be determined using haplotype deconvolution. The increase in the contribution of the haplotype to the transplanted organ would be useful to indicate the increased contribution of the organ to the plasma DNA. In various embodiments, the baseline level can be determined from a cohort of transplant recipients not having rejection or from a cohort of transplant recipients having rejection. When using recipients having rejection, the baseline level can be determined as below those from a cohort of transplant recipients having rejection.

[0266] As mentioned above, the donor may have two identical haplotypes or the recipient can also have two identical haplotypes. Furthermore, the donor and recipient may share a haplotype. As long as the donor or the recipient has a unique haplotype, a change in a percentage of cell-free DNA molecules from the donor tissue can be determined. In the former, a rejection will be detected when one sees an increase in the contribution of the donor-unique haplotype in plasma (or other sample). In the latter, a rejection will be detected when one sees a decrease in the contribution of the recipient-unique haplotype in plasma.

[0267] Accordingly, some embodiments can use a first haplotype that is present in normal cells of the organism and not being present in abnormal cells that may be in the mixture. This would correspond to the latter example above, when the recipient has a unique haplotype. Another

example is when a patient has a unique haplotype in healthy cells compared to a tumor (e.g., previously found in the organism). In this embodiment, the first tissue type can be determined to have the disease state when the first separation value is less than the threshold value.

[0268] In some embodiments, if the transplanted organ is detected as being rejected, treatment can be provided. For example, a change is dosage of anti-rejection medication can be provided. As another example, a new organ can be obtained, and surgery can be performed to remove the old transplanted organ and put in the new transplanted organ.

B. Hepatocellular carcinoma (HCC)

[0269] As an example for determining a tissue of origin for a cancer signature or aberration (or monitoring for a tumor that was known to exist or have existed), we analyzed the plasma of an HCC patient. The tumor and the blood cells of the patients were sequenced to identify the cancer-specific single nucleotide mutations. Sequenced DNA fragments carrying a cancer-specific mutation and at least one CpG site were used for downstream methylation deconvolution analysis. A total of 11,968 fragments were used for deconvolution analysis. In addition to the methylation profiles from the normal tissue organs, we have also included the methylation profile of HCC tissues as candidate tissue of origin.

[0270] In another embodiment, more types of tumor tissues can be considered as candidate tissues for the mutations. In one embodiment, the methylation profiles of the common cancers, for example but not limited to colorectal cancer, lung cancer, breast cancer, pancreatic cancer, prostate cancer, bladder cancer, cervical cancer and ovarian cancer can be included as candidate tissues. In yet another embodiment, only the most possible cancers specific to the patient can be included in the analysis. For example, in female patients, the breast cancer, ovarian cancer, colorectal cancer and cervical cancer are considered. In yet another embodiment, the ethnic origin and the age are considered in the selection of the candidate tissues.

[0271] Table 7 shows methylation deconvolution of the plasma DNA fragments carrying cancer-associated mutations. The deconvolution analysis accurately determined that the DNA fragments carrying cancer-associated mutations are predominantly derived from the liver cancer tissues.

Tissue	Contribution (%)

Liver	0.0
Lung	0.0
Colon	0.0
Small intestines	0.0
Pancreas	0.0
Adrenal glands	0.0
Esophagus	0.0
Adipose tissues	0.0
Heart	0.0
Brain	0.0
T-cells	0.0
B-cells	0.0
Neutrophil	4.6
Liver cancer	95.4
Placenta	0.0

Table 7. Fractional contributions for HCC patient using cancer mutation.

[0272] In some embodiments, the tumor can initially be identified by detecting a copy number aberration, e.g., as described in U.S. Patent Nos. 8,741,811 and 9,121,069. The particular tissue of origin can be determined, e.g., as described in U.S. Patent Application 14/994,053 based on patterns of copy number aberrations previously identified in various tumors. Once the tumor has been identified, treatment can be performed, e.g., by surgery, radiotherapy, or chemotherapy. Either way, a biopsy can be obtained after the tissue of origin is determined. A cancer-specific point mutation can be determined from the biopsy or from DNA fragments in plasma (e.g., as described in U.S. patent Publication 2014/0100121, or other mixtures that are associated with the copy number aberration.

[0273] After treatment, a key change would be the disappearance of the genomic aberrations, including the copy number aberration and point mutation. When these aberrations are gone, the analysis of the genomic signature of the point mutation in the affected regions would give a change in the tissue contribution via the methylation deconvolution analysis. If the tumor comes back in the future, the cancer-associated changes in tissue composition (as determined using methylation deconvolution analysis) would be seen again. For example, the fractional contribution can be compared to a reference fractional contribution, and if a change is detected, then new courses of treatment can be provided.

[0274] In various embodiments, the cancer-specific mutation can be on only one haplotype or on both haplotypes, e.g., in a manner similar to the donor example above. Thus, as with the donor, the fractional contributions can be determined for each of Hap I and Hap II separately, if different signatures exist; such different signatures can exist at different loci. In other embodiments, a single fractional contribution can be determined for both haplotypes, e.g., when they share a signature.

C. Imprinting

[0275] In another embodiment, the haplotype deconvolution analysis can be applied for the analysis of the genomic regions showing tissue-specific imprinting. It has been shown that the differential methylation of the paternally and maternally inherited alleles in different tissue organs is a common phenomenon (Baran et al. Genome Res 2015;25:927-36). Haplotype deconvolution would be useful for the monitoring of the contribution of the organ exhibiting tissue-specific imprinting. For example, when the paternally and maternally inherited haplotypes have different methylation status in the liver but not in other tissues, methylation deconvolution can be performed on both the paternally and maternally inherited haplotypes. In one embodiment, both the paternal and maternal methylation patterns can be included as candidate tissues in the analysis.

D. Method using Genomic Signature

[0276] FIG. 19 is a flowchart illustrating a method 1900 of analyzing a biological sample of an organism to detect whether a first tissue type has a disease state associated with a first haplotype according to embodiments of the present invention. The biological sample includes a mixture of cell-free DNA molecules from a plurality of tissues types, including a first tissue type. Method 1900 is performed at least partially using a computer system.

[0277] At block 1910, a plurality of cell-free DNA molecules from the biological sample are analyzed. Block 1910 can be performed using techniques described in block 140 of method 100 of FIG. 1. For example, at least 1,000 cell-free DNA molecules can be analyzed to determine where the cell-free DNA molecules are located, and methylation levels can be measured as described below. Further, the cell-free DNA molecules are analyzed to determine a respective allele of the cell-free DNA molecule. For example, an allele of a DNA molecule can be determined from a sequence read or from a particular probe that hybridizes to the DNA molecule.

[0278] At block 1920, one or more loci are identified. Each locus has a first allele on a first haplotype of a first chromosomal region. The first haplotype has a property of either: (1) not being present in healthy cells of the organism, but instead may be from a tumor or transplanted tissue, as examples; or (2) being present in normal cells of the organism and not being present in abnormal cells that may be in the mixture. Thus, the first haplotype has a genomic signature. In this manner, there is a difference between the healthy (normal) cells and the abnormal cells, thereby allowing embodiments to track a fractional contribution of one or the other, or both, so as to track an extent (e.g., fractional contribution) of the abnormal cells. With property (1), the first haplotype is associated with a disease state, e.g., cancer or a rejection of transplanted tissue. Thus, a particular cancer can have the first haplotype in a cancer genome of the particular cancer.

[0279] The one or more first alleles may be identified at the one or more loci on the first haplotype by obtaining a tissue sample (e.g., of the tumor or transplanted tissue) and analyzing DNA molecules of the tissue sample to determine the first haplotype. Such a tissue sample may be obtained from a biopsy, and method 1900 can be used to test if the cancer has metastasized to other tissues, or has recurred after surgery. Each one of the loci may be a heterozygous locus or a homozygous locus in the abnormal cells. For example, in FIG. 18, locus 1 and locus 3 are homozygous in the donor organ. But, ultimately, more than one allele would be observed in the plasma for all the loci, as each locus would have a signature for healthy cells or for abnormal cells. Thus, two haplotypes would exist across the tissue types, but a single tissue type might have only one haplotype in a region being analyzed.

[0280] At block 1930, a first set of the plurality of cell-free DNA molecules is identified. Each of the plurality of cell-free DNA molecules is located at any one of the loci from block 1920 and includes a corresponding first allele at the one locus, so that the cell-free DNA molecule can be identified as corresponding to the first haplotype. Each of the first set of cell-free DNA molecules also includes at least one of N genomic sites, where the genomic sites are used to measure the methylation levels. N is an integer, e.g., greater than or equal to 2, 3, 4, 5, 10, 20, 50, 100, 200, 500, 1,000, 2,000, or 5,000.

[0281] At block 1940, N first mixture methylation levels are measured at the N genomic sites using the first set of the plurality of cell-free DNA molecules. One first mixture methylation level can be measured for each of the N genomic sites. Block 1940 can be performed in a similar

manner as block 150 of method 100 of FIG. 1. In some embodiments, the measurement of the methylation level of a DNA molecule can use methylation-aware sequencing results, which may also be used to determine the location and respective allele of the DNA molecule.

[0282] At block 1950, a first fractional contribution of the first tissue type in the mixture is determined using the N first methylation levels. In some embodiments, block 1950 can be performed via blocks 160 and 170 of method 100 of FIG. 1. Thus, a fractional contribution can be determined simultaneously for a panel of M tissue types. Block 1950 may use N issue-specific methylation levels at N genomic sites, determined for each of M tissue types, e.g., as in block 120 of method 100 of FIG. 1.

[0283] At block 1960, a separation value between the first fractional contribution and a reference fractional contribution is computed. Examples of separation values are described herein. The reference fractional contribution can be determined using samples from organisms that are healthy for the first tissue type. For a transplant example, the reference fractional contribution can be determined from one or more measurements of biological samples of organisms whose transplanted first tissue is not being rejected.

[0284] At block 1970, the separation value can be compared to a threshold value to determine a classification of whether the first tissue type has a disease state. For example, if the first haplotype is associated with cancer, then an appreciable first fractional contribution indicates that the first tissue type has cancer, as can be measured by the separation value exceeding the threshold value (e.g., when the reference fractional contribution is zero). The amount that the first fractional contribution exceeds the threshold value can indicate a certain level of cancer. As another example, the first haplotype can be specific to transplanted tissue, and a high contribution relative to the reference can indicate the organism is rejecting the transplanted tissue.

[0285] In an embodiment where the first haplotype is present in normal cells of the organism and not present in abnormal cells that may be in the mixture, the first tissue type can be determined to have the disease state when the first separation value is less than the threshold value. An example of a disease state is preeclampsia which can be associated with a spectrum of pathologic changes in a fetal tissue such as the placenta. As an example, in such a situation, if the first haplotype is specific to the fetus, e.g. paternally inherited haplotype, it may be increased in the maternal plasma in a pregnancy complicated with preeclampsia.

[0286] In some embodiments, a second haplotype for the diseased tissue, e.g., the transplanted tissue or a tumor may also be used. Thus, a second fractional contribution can be computed and compared to the reference fractional contribution. Accordingly, a second set of the plurality of cell-free DNA molecules can each be located at any one of the one or more loci, include a corresponding second allele on a second haplotype of the first chromosomal region, and include at least one of the N genomic sites. The second haplotype would have the same property of being only from healthy cells or abnormal cells.

[0287] A plurality of tissue types can be tested (e.g., using method 100 of FIG. 1), so as to determine a tissue of origin of the first haplotype, e.g., when it is associated with cancer. Accordingly, fractional contributions of other tissue types in the mixture can be determined using the N first methylation levels, and corresponding separation values between the corresponding fractional contributions and respective reference fractional contribution can be compared to the threshold value to determine a classification of whether each of the other tissue types has the particular cancer. Different tissue could have different reference fractional contributions.

VIII. IDENTIFYING TISSUE OF ORIGIN OF CNA OF CANCER

[0288] In some embodiments, an origin of a tumor may not be known. Thus, it can be difficult to identify point mutations in a tumor, as may be used for method 1900 of FIG. 19 or other methods described herein. Additionally, a tumor may not have a significant number of point mutations, but may have chromosomal regions exhibiting amplifications and deletions (examples of copy number aberrations).

[0289] To address this problem, embodiments can use a copy number analysis to identify regions that exhibit a copy number aberration (CNA). Typically, a CNA occurs on only one haplotype of a region. As only one haplotype has an amplification or a deletion, there will be a relatively large difference between the fractional contributions of the tissue type within which the tumor resides.

[0290] The CNA analysis may be performed in a variety of ways, e.g., as described in U.S. Patent Nos. 8,741,811 and 9,121,069. For example, the human genome (or genome for other type of organism) can be partitioned into approximately 3,000 non-overlapping 1-Mb bins. The number of reads mapping to each 1-Mb bin can be determined. After correcting for GC bias (Chen EZ, et

al. (2011) *PLoS One* 6(7):e21791), the sequence read density of each bin can be calculated. For each bin, the sequence read density of the test case can be compared to the values of the reference control subjects. Copy number gains and losses may be defined as 3 standard deviations above and below, respectively, the mean of the controls. Accordingly, identifying a first chromosomal region as exhibiting a copy number aberration can be based on a first amount of cell-free DNA molecules that are located in the first chromosomal region.

[0291] To determine the tissue origin of copy number aberrations in plasma, plasma DNA tissue mapping can be performed using the methylation markers located within the genomic regions exhibiting such aberrations in plasma. In the examples below for the cancer patients, mapping of plasma DNA copy number aberrations was performed only in cases with aberrations affecting a contiguous chromosome region of at least 30 Mb so that a sufficient number of methylation markers could be used for mapping.

A. Identifying regions with copy number aberration (CNA)

[0292] A 62-year-old male patient with HCC was recruited from the Department of Surgery, Prince of Wales Hospital, Hong Kong with informed consent. Ten milliliters of venous blood were collected in EDTA tubes at diagnosis and 3 months after the resection of the tumor. The blood samples were centrifuged at 3000 g for 10 minutes to separate the blood cells from the plasma. The plasma was recentrifuged at 30000 g for 10 minutes to remove the remaining cells.

[0293] DNA extracted from the blood cells was used for phasing the SNPs to construct the haplotypes of the patient using the 10x genomics platform following the manufacturer's instruction. High molecular weight DNA was extracted from blood or tissue samples using MagAttract HMW DNA kit (QIagen, Germany). The quality of DNA was verified by Genomic DNA Analysis ScreenTape on a 4200 TapeStation system (Agilent, Germany). DNA was quantified by dsDNA HS Assay kit on a Qubit 3.0 fluorometer (Thermo Fisher Scientific, Waltham, MA). Sample indexing and library preparation were performed using the GemCode system and its associated reagents (10X Genomics, Pleasanton, CA) (Zheng et al. Nat Biotechnol. 2016 Mar;34:303-11). In brief, 1ng of DNA was inputted for GEM reactions in which individual DNA molecules were partitioned to introduce specific barcodes and extend the DNA. After GEM reactions, sequencing libraries were prepared according to the manufacturer's recommendations.

The libraries were quantified by qPCR using KAPA Library Quantification Kit (KAPA Biosystems, Wilmington, MA). The normalized libraries were sequenced on a HiSeq 2500 sequencer (Illumina, San Diego, CA), with paired-end sequencing of 98-bp, 14-bp I5 and 8-bp I7 index reads. Sequencing results were analyzed using Long Ranger software suite (10X Genomics) so that all heterozygous SNPs were phased and the two haplotypes of the patient was determined.

[0294] The plasma samples were sequenced using the Illumina to a depth of 17x. Copy number aberrations were detected in the plasma of the HCC patient according to the method as previously described (Chan et al. Clin Chem. 2013;59:211-24).

[0295] FIG. 20 shows a plot of copy number aberrations detected in the plasma of an HCC patient according to embodiments of the present invention. The inner circle represents the result of the plasma sample collected at diagnosis (pre-operation) and the outer circle represents the result of the plasma sample collected at 3 months after the resection of the tumor (post-operation). Each dot represents a 1-Mb region. The green, red, and grey dots represent regions with copy number gain, copy number loss and no copy number change, respectively. Copy number aberrations were detected in the plasma sample at diagnosis and these changes disappeared after the tumor was removed.

[0296] In FIG. 20, two regions are highlighted for having a CNA. Region 2010 has a copy number gain, and region 2020 has a copy number loss. The haplotypes of these regions can be determined using any tissue sample of the subject, and not just a tumor sample. The difference in copy number is what is driving the difference in the fractional contributions, and that difference should be greatest in the tissue type with the tumor.

B. Determining the tissue origin of the copy number aberrations

[0297] We performed methylation deconvolution analysis for the two haplotypes independently. For illustration purposes, the two haplotypes are named Hap I and Hap II. The plasma DNA molecules covering heterozygous SNPs and at least one CpG site were used for this analysis. Plasma DNA molecules carrying the SNP alleles on Hap I were analyzed independently from those carrying alleles on Hap II. The methylation status of the CpG sites were used for methylation deconvolution for molecules mapped to Hap I and Hap II independently. As a result, the tissue contribution to Hap I and Hap II in plasma DNA could be determined.

[0298] First we focused on the regions with amplification. For illustration purposes, we analyzed the amplified region on chromosome 1q as an example.

	At diagnosis	After tumor resection
Hap I	34,119	11,131
Hap II	26,582	11,176

[0299] Table 11 shows the number of sequence reads from the two haplotypes. At diagnosis, the number of reads mapped to Hap I was increased compared with the number of reads mapped to Hap II. This indicated that Hap I is amplified relative to Hap II. This observation is compatible with the fact that a particular chromosome is duplicated in cancer rather than both homologous chromosomes being amplified to the same extent, which was in line with the fact that the copy number aberrations occurs preferentially on one haplotype (Adey A. et al, Nature. 2013;500:207-11;LaFramboise T. et al, PLoS Comput Biol. 2005;1(6):e65). The difference in the dosage of the two haplotypes disappeared after the tumor was resected. The difference in the absolute sequence reads number between the plasma samples taken at diagnosis and after tumor resection was due to the difference in the total number of sequence reads generated for the two plasma samples.

	At diagnosis			After tumor resection		
	Hap I	Hap II	Difference	Hap I	Hap II	Difference
liver	19.7	8.0	11.7	21.3	21.9	-0.6
Lung	5.4	0	5.4	0	0	0
Colon	0	0	0	0	0	0
Brain	0	0	0	9.0	9.0	0
Heart	0	17.0	-17	3.0	2.5	-0.5
Blood cell	74.9	75.0	0	66.7	66.6	0.1
Total	100	100	0	100	100	0

[0300] Table 12 shows a percentage contribution of different tissues to plasma DNA for the two haplotypes at diagnosis and after tumor resection. At diagnosis, the contributions of liver to plasma DNA were 19.7% and 8.0% for Hap I and Hap II, respectively. A difference of 11.7% was the highest among different types of tissue. This indicated that the dosage difference between Hap I and Hap II in the plasma was most likely contributed from the contribution of the liver. This further indicated that the likely origin of the chromosomal aberration was from the liver because the copy number changes was most likely due to the duplication of Hap I in the sequence read

counts analysis. In another embodiment, the difference in the contribution for Hap I and Hap II can be ranked to indicate the relative likelihood of different tissues being the source of the copy number aberrations.

[0301] The value for the heart is -17, which is in the opposite direction of the copy number aberration identified by Table 11. Thus, although the absolute value for the heart is larger than the absolute value for the liver, the opposite sign would discount the heart as being a viable candidate for the tissue type of the origin of the tumor. As the total contribution of all organs is 100%, the positive difference in the contributions of liver results in other tissues having negative values.

[0302] Similarly, this haplotype-specific methylation deconvolution can also be performed on regions with copy number loss. For illustration purpose, we performed this analysis on a region on chromosome 1p that exhibited copy number loss.

	At diagnosis	After tumor resection
Hap I	19,973	8,323
Hap II	12,383	7,724

[0303] Table 13 shows a number of sequence reads from the two haplotypes. At diagnosis, the number of reads mapped to Hap II was decreased compared with the number of reads mapped to Hap I. In tumor tissues, most of the regions with chromosome copy number loss would only involve the deletion of one of the two chromosomes. Thus the relative reduction in the dosage of Hap II was compatible with the deletion of Hap II. The difference in the dosage of the two haplotypes disappeared after the tumor was resected indicated that the amount of tumor-derived DNA had decreased or disappeared from the plasma.

	At diagnosis			After tumor resection		
	Hap I	Hap II	Difference (Hap I – Hap II)	Hap I	Hap II	Difference (Hap I – Hap II)
liver	13.3	5.5	7.8	10.2	13.2	-3
Lung	0	0	0	4.1	0.5	3.6
Colon	3.8	0	3.8	8.6	17.5	-8.9
Brain	0	0	0	0	0	0
Heart	3.7	0	3.7	25.5	19.4	6.1
Blood cell	79.2	94.5	-15.3	51.6	49.4	2.2
Total	100	100	0	100	100	0

[0304] Table 14 shows a percentage contribution of different tissues to plasma DNA for the two haplotypes at diagnosis and after tumor resection. At diagnosis, the contributions of liver to plasma DNA were 13.3% and 5.5% for Hap I and Hap II, respectively. A difference of 7.8% was the highest among different types of tissue. This indicated that the dosage difference between Hap I and Hap II in the plasma was most likely contributed from the contribution of the liver. This further indicated that the likely origin of the chromosomal aberration was from the liver because the copy number changes was most likely due to the deletion of Hap II in the sequence read counts analysis. In another embodiment, the difference in the contribution for Hap I and Hap II can be ranked to indicate the relative likelihood of different tissues being the source of the copy number aberrations.

C. Method of determining tissue origin of tumor

[0305] FIG. 21 is a flowchart illustrating a method of analyzing a biological sample of an organism to identify an origin of a chromosomal aberration according to embodiments of the present invention. The biological sample includes a mixture of cell-free DNA molecules from a plurality of tissues types that include a first tissue type.

[0306] At block 2110, a plurality of cell-free DNA molecules from the biological sample are analyzed. Block 2110 can be performed using techniques described in block 1910 of FIG. 1 and block 140 of method 100 of FIG. 1, as well as other blocks describing similar features.

[0307] At block 2115, a first chromosomal region is identified as exhibiting a copy number aberration in the organism based on a first amount of cell-free DNA molecules that are located in the first chromosomal region. As an example, plasma DNA analysis is performed to identify regions that exhibit copy number aberrations. The aberration can correspond to over or under-representation. In some embodiments, the genome can be separated into bins (e.g., 1-Mb bins), and the amount of cell-free DNA molecules from a particular bin can be determined (e.g., by mapping sequence reads to that part of a reference genome). The amount for a particular bin can be normalized (e.g., with respect to an average amount for a bin) can an over or under-representation can be identified.

[0308] Other techniques than counting DNA molecules mapping to a particular region can be used. For example, a distribution of sizes of the DNA molecules aligning to the first chromosomal

region can be used to detect the CNA. For example, cell-free tumor DNA is smaller than cell-free DNA from normal cells. This difference in size can be used to detect differences in a size distribution (e.g., average size or ratio of number of DNA molecules at different sizes) between two haplotypes for the region, or between the region and another region.

[0309] At block 2120, a first haplotype and a second haplotype of the organism in the first chromosomal region are determined. The two haplotypes may have been determined as part of block 2115. The two haplotypes can be determined using the same cell-free mixture or from a different sample, e.g., a cellular sample.

[0310] At block 2130, one or more heterozygous loci of the first chromosomal region are identified. Each heterozygous locus includes a corresponding first allele in the first haplotype and a corresponding second allele in the second haplotype. Block 2130 may be performed in a similar manner as other similar blocks of methods described herein.

[0311] At block 2140, a first set of the plurality of cell-free DNA molecules are identified. Each DNA molecule of the first set is located at any one of the one or more heterozygous loci, includes the corresponding first allele of the heterozygous locus, and includes at least one of N genomic sites. N is an integer greater than or equal to 2. Block 2140 may be performed in a similar manner as other similar blocks of methods described herein.

[0312] At block 2150, N first mixture methylation levels at the N genomic sites are measured using the first set of the plurality of cell-free DNA molecules. Block 2150 may be performed in a similar manner as other similar blocks of methods described herein.

[0313] At block 2160, a second set of the plurality of cell-free DNA molecules is identified. Each DNA molecule of the second set is located at any one of the one or more heterozygous loci, includes the corresponding second allele of the heterozygous locus, and includes at least one of the N genomic sites. Block 2160 may be performed in a similar manner as other similar blocks of methods described herein.

[0314] In some embodiments, a first number of cell-free DNA molecules in the first set of the plurality of cell-free DNA molecules can be determined, and a second number of cell-free DNA molecules in the second set of the plurality of cell-free DNA molecules can be determined, e.g., as shown in table 11. It can be determined which number is higher, thereby providing information

about the expected separation value for the tissue of origin, e.g., which haplotype should have a higher fractional contribution.

[0315] The first set of the plurality of cell-free DNA molecules can have a first size distribution, and the second set of the plurality of cell-free DNA molecules can have a second size distribution. A statistical value of a size distribution of the DNA molecules can be determined for each haplotype, thereby providing a first statistical value and a second statistical value. The haplotype with the smaller size distribution would be expected to have a higher copy number than the other haplotype, as tumor cell-free DNA is known to be smaller, as described in U.S. Patent No. 8,741,811. Examples of a statistical value of a size distribution are a ratio of a number of DNA molecules at different sizes, an average size, or a percentage of the DNA molecules at a specific size (e.g., below a size cutoff).

[0316] At block 2170, N second mixture methylation levels at the N genomic sites are measured using the second set of the plurality of cell-free DNA molecules. Block 2170 may be performed in a similar manner as other similar blocks of methods described herein.

[0317] Blocks 2180 and 2190 may be performed for each of a plurality of M tissue types. The M tissue types can include a default list of tissue types that are screened and for which reference methylation levels may be known. The default list can include tissues that cancer is most predominantly seen. M is an integer greater than one.

[0318] At block 2180, a computer system determines a corresponding first fractional contribution of the tissue type in the mixture using the N first methylation levels. The computer system determines a corresponding second fractional contribution of the tissue type in the mixture using the N second methylation levels. Block 2180 may be performed in a similar manner as other similar blocks of methods described herein.

[0319] At block 2190, a corresponding separation value between the corresponding first fractional contribution and the corresponding second fractional contribution is computed. Various separation values can be used, e.g., as are described herein.

[0320] At block 2195, the first tissue type is identified as being an origin of the copy number aberration based on a first separation value of the first tissue type having a maximal value among the corresponding separate values. The determination can require that the highest separation value

be sufficiently higher than the second highest separation value. For example, the difference can be required to be at least a threshold value, e.g., 1%, 2%, 35, 4%, 5%, 6%, or 7%. In one implementation, a difference between the first separation value and a next highest separation value can be compared to a threshold to determine a classification of how likely the first tissue type is the origin of the copy number aberration. Thus, even if the difference is not above the threshold, a probability or other classification can be provided. For instance, a linear relationship can be used from 0 to the threshold, where the probability is 100% once the difference is equal to the threshold.

[0321] Depending on how the separation value is determined, the maximal value can be a maximum negative number or a maximum positive number. For example, the difference values in table 14 could be determined using Hap II – Hap I. Whether the maximum should be a positive or a negative value can be determined using the analysis of DNA molecules on each haplotype, e.g., a count as in Table 13 or a size analysis as described above. In some implementations, the separation value can always be determined such that a maximum positive value is expected, e.g., by subtracting the fractional contribution of the haplotype with a lower copy number from the fractional contribution of the haplotype with the higher copy number.

[0322] After the origin is identified, an investigation using imaging modalities, e.g. computed tomography (CT) scan or magnetic resonance imaging (MRI), of the subject (entire subject or specifically of the candidate organ) can be performed to confirm or rule out the presence of a tumor in the organ. If presence of a tumor is confirmed, treatment can be performed, e.g., surgery (by a knife or by radiation) or chemotherapy.

IX. COMPUTER SYSTEM

[0323] Any of the computer systems mentioned herein may utilize any suitable number of subsystems. Examples of such subsystems are shown in FIG. 22 in computer apparatus 10. In some embodiments, a computer system includes a single computer apparatus, where the subsystems can be the components of the computer apparatus. In other embodiments, a computer system can include multiple computer apparatuses, each being a subsystem, with internal components. A computer system can include desktop and laptop computers, tablets, mobile phones and other mobile devices.

[0324] The subsystems shown in FIG. 22 are interconnected via a system bus 75. Additional subsystems such as a printer 74, keyboard 78, storage device(s) 79, monitor 76, which is coupled to display adapter 82, and others are shown. Peripherals and input/output (I/O) devices, which couple to I/O controller 71, can be connected to the computer system by any number of means known in the art such as input/output (I/O) port 77 (e.g., USB, FireWire®). For example, I/O port 77 or external interface 81 (e.g. Ethernet, Wi-Fi, etc.) can be used to connect computer system 10 to a wide area network such as the Internet, a mouse input device, or a scanner. The interconnection via system bus 75 allows the central processor 73 to communicate with each subsystem and to control the execution of a plurality of instructions from system memory 72 or the storage device(s) 79 (e.g., a fixed disk, such as a hard drive, or optical disk), as well as the exchange of information between subsystems. The system memory 72 and/or the storage device(s) 79 may embody a computer readable medium. Another subsystem is a data collection device 85, such as a camera, microphone, accelerometer, and the like. Any of the data mentioned herein can be output from one component to another component and can be output to the user.

[0325] A computer system can include a plurality of the same components or subsystems, e.g., connected together by external interface 81 or by an internal interface. In some embodiments, computer systems, subsystem, or apparatuses can communicate over a network. In such instances, one computer can be considered a client and another computer a server, where each can be part of a same computer system. A client and a server can each include multiple systems, subsystems, or components.

[0326] Aspects of embodiments can be implemented in the form of control logic using hardware (e.g. an application specific integrated circuit or field programmable gate array) and/or using computer software with a generally programmable processor in a modular or integrated manner. As used herein, a processor includes a single-core processor, multi-core processor on a same integrated chip, or multiple processing units on a single circuit board or networked. Based on the disclosure and teachings provided herein, a person of ordinary skill in the art will know and appreciate other ways and/or methods to implement embodiments of the present invention using hardware and a combination of hardware and software.

[0327] Any of the software components or functions described in this application may be implemented as software code to be executed by a processor using any suitable computer language

such as, for example, Java, C, C++, C#, Objective-C, Swift, or scripting language such as Perl or Python using, for example, conventional or object-oriented techniques. The software code may be stored as a series of instructions or commands on a computer readable medium for storage and/or transmission. A suitable non-transitory computer readable medium can include random access memory (RAM), a read only memory (ROM), a magnetic medium such as a hard-drive or a floppy disk, or an optical medium such as a compact disk (CD) or DVD (digital versatile disk), flash memory, and the like. The computer readable medium may be any combination of such storage or transmission devices.

[0328] Such programs may also be encoded and transmitted using carrier signals adapted for transmission via wired, optical, and/or wireless networks conforming to a variety of protocols, including the Internet. As such, a computer readable medium may be created using a data signal encoded with such programs. Computer readable media encoded with the program code may be packaged with a compatible device or provided separately from other devices (e.g., via Internet download). Any such computer readable medium may reside on or within a single computer product (e.g. a hard drive, a CD, or an entire computer system), and may be present on or within different computer products within a system or network. A computer system may include a monitor, printer, or other suitable display for providing any of the results mentioned herein to a user.

[0329] Any of the methods described herein may be totally or partially performed with a computer system including one or more processors, which can be configured to perform the steps. Thus, embodiments can be directed to computer systems configured to perform the steps of any of the methods described herein, potentially with different components performing a respective steps or a respective group of steps. Although presented as numbered steps, steps of methods herein can be performed at a same time or in a different order. Additionally, portions of these steps may be used with portions of other steps from other methods. Also, all or portions of a step may be optional. Additionally, any of the steps of any of the methods can be performed with modules, units, circuits, or other means for performing these steps.

[0330] The specific details of particular embodiments may be combined in any suitable manner without departing from the spirit and scope of embodiments of the invention. However, other

embodiments of the invention may be directed to specific embodiments relating to each individual aspect, or specific combinations of these individual aspects.

[0331] The above description of example embodiments of the invention has been presented for the purposes of illustration and description. It is not intended to be exhaustive or to limit the invention to the precise form described, and many modifications and variations are possible in light of the teaching above.

[0332] A recitation of "a", "an" or "the" is intended to mean "one or more" unless specifically indicated to the contrary. The use of "or" is intended to mean an "inclusive or," and not an "exclusive or" unless specifically indicated to the contrary. Reference to a "first" component does not necessarily require that a second component be provided. Moreover reference to a "first" or a "second" component does not limit the referenced component to a particular location unless expressly stated.

[0333] All patents, patent applications, publications, and descriptions mentioned herein are incorporated by reference in their entirety for all purposes. None is admitted to be prior art.

WHAT IS CLAIMED IS:

1. A method of determining a portion of a fetal genome of an unborn fetus of a pregnant female using a biological sample from the pregnant female, wherein the biological sample including a mixture of cell-free DNA molecules from a plurality of tissues types, including maternal tissue types and a fetal tissue type, the fetus having a father and a mother being the pregnant female, the method comprising:

analyzing, by a computer system, a plurality of cell-free DNA molecules from the biological sample, the plurality of cell-free DNA molecules being at least 1,000 cell-free DNA molecules, wherein analyzing a cell-free DNA molecule includes:

identifying a location of the cell-free DNA molecule in a reference human genome; and

determining a respective allele of the cell-free DNA molecule;

determining a first haplotype and a second haplotype of a first chromosomal region of a first parental genome of a first parent of the fetus;

identifying one or more heterozygous loci of the first chromosomal region of the first parental genome, each heterozygous locus including a corresponding first allele in the first haplotype and a corresponding second allele in the second haplotype;

identifying a first set of the plurality of cell-free DNA molecules that each:

is located at any one of the one or more heterozygous loci,

includes the corresponding first allele of the heterozygous locus, and

includes at least one of N genomic sites, N being an integer greater than or equal to 2;

measuring N first mixture methylation levels at the N genomic sites using the first set of the plurality of cell-free DNA molecules;

determining, by the computer system, a first fractional contribution of the fetal tissue type in the mixture using the N first methylation levels;

identifying a second set of the plurality of cell-free DNA molecules that each:

is located at any one of the one or more heterozygous loci,

includes the corresponding second allele, and

includes at least one of the N genomic sites;

measuring N second mixture methylation levels at the N genomic sites using the second set of the plurality of cell-free DNA molecules;

determining, by the computer system, a second fractional contribution of the fetal tissue type in the mixture using the N second methylation levels;

computing a first separation value between the first fractional contribution and the second fractional contribution;

determining the portion of fetal genome at the one or more heterozygous loci based on the first separation value.

2. The method of claim 1, wherein the one or more heterozygous loci are a first plurality of heterozygous loci.

3. The method of claim 2, wherein the first haplotype is a first maternal haplotype, and wherein the second haplotype is a second maternal haplotype.

4. The method of claim 3, further comprising:

identifying a second plurality of heterozygous loci of the first chromosomal region in a paternal genome, each of the second plurality of heterozygous loci including a corresponding third allele in a first paternal haplotype and a corresponding fourth allele in a second paternal haplotype, wherein the paternal genome corresponds to the father of the fetus;

identifying a third set of the plurality of cell-free DNA molecules that each:

is located at any one of the second plurality of heterozygous loci,

includes the corresponding third allele of the heterozygous locus, and

includes at least one of K genomic sites;

measuring K third mixture methylation levels at the K genomic sites using the third set of the second plurality of cell-free DNA molecules;

determining a third fractional contribution of the fetal tissue type in the mixture using the K third methylation levels;

identifying a fourth set of the plurality of cell-free DNA molecules that each:

is located at any one of the second plurality of heterozygous loci,

includes the corresponding fourth allele of the heterozygous locus, and

includes at least one of the K genomic sites;

measuring K fourth mixture methylation levels at the K genomic sites using the fourth set of the second plurality of cell-free DNA molecules;

determining a fourth fractional contribution of the fetal tissue type in the mixture using the K fourth methylation levels;

computing a second separation value between the third fractional contribution and the fourth fractional contribution;

determining the portion of fetal genome at the second plurality of heterozygous loci based on the second separation value.

5. The method of claim 2, wherein the first separation value is a ratio of the first fractional contribution and the second fractional contribution, and wherein the portion of the fetal genome is determined to have the first haplotype and the second haplotype when the ratio is equal to one within a cutoff value.

6. The method of claim 2, wherein the first separation value is a difference of the first fractional contribution and the second fractional contribution.

7. The method of claim 2, wherein the portion of the fetal genome is determined to have one or more copies of the first haplotype and no copies of the second haplotype when the first separation value is greater than a threshold value.

8. The method of claim 2, wherein the portion of the fetal genome is determined to have one or more copies of the second haplotype and no copies of the first haplotype when the first separation value is less than a threshold value.

9. The method of claim 1, wherein N is 10 or more.

10. The method of claim 1, wherein the N mixture methylation levels form a methylation vector b, and wherein determining the first fractional contribution of the fetal tissue type includes:

for each of M tissue types:

obtaining N tissue-specific methylation levels at the N genomic sites, N being greater than or equal to M, wherein the tissue-specific methylation levels form a matrix A of dimensions N by M, the M tissue types including the fetal tissue type;

solving for a composition vector x that provides the methylation vector b for the matrix A; and

for each of one or more components of the composition vector x:

using the component to determine a corresponding fractional contribution of a corresponding tissue type of the M tissue types in the mixture.

11. The method of claim 10, wherein the M tissue types include liver, lung, neutrophils, lymphocytes, erythroblasts, heart, large intestines, small intestines, and placenta, wherein the fetal tissue type is the placenta.

12. The method of claim 1, wherein measuring N first mixture methylation levels at the N genomic sites includes analyzing methylation-aware sequencing results, and wherein the locations of the plurality of cell-free DNA molecules are determined using the methylation-aware sequencing results.

13. The method of claim 1, wherein the first separation value includes a ratio of the first fractional contribution and the second fractional contribution.

14. The method of claim 1, wherein the first separation value includes a difference of the first fractional contribution and the second fractional contribution.

15. A method of determining a portion of a fetal genome of an unborn fetus of a pregnant female using a biological sample from the pregnant female, wherein the biological sample including a mixture of cell-free DNA molecules from a plurality of tissues types, including maternal tissue types and a fetal tissue type, the fetus having a father and a mother being the pregnant female, the method comprising:

analyzing, by a computer system, a plurality of cell-free DNA molecules from the biological sample, the plurality of cell-free DNA molecules being at least 1,000 cell-free DNA molecules, wherein analyzing a cell-free DNA molecule includes:

identifying a location of the cell-free DNA molecule in a reference human genome; and

determining a respective allele of the cell-free DNA molecule;

determining a first haplotype of a first chromosomal region of a first parental genome of a first parent of the fetus;

identifying one or more heterozygous loci of the first chromosomal region of the first parental genome, each heterozygous locus including a corresponding first allele in the first haplotype and a corresponding second allele in a second haplotype of the first chromosomal region of the first parental genome;

identifying a first set of the plurality of cell-free DNA molecules that each:

is located at any one of the one or more heterozygous loci,

includes the corresponding first allele of the heterozygous locus, and

includes at least one of N genomic sites, N being an integer greater than or equal to 2;

measuring N first mixture methylation levels at the N genomic sites using the first set of the plurality of cell-free DNA molecules;

determining, by the computer system, a first fractional contribution of the fetal tissue type in the mixture using the N first methylation levels; and

comparing the first fractional contribution to a reference value to determine whether the fetus inherited the first haplotype at the first chromosomal region, wherein the fetus inherited the first haplotype at the first chromosomal region when the first fractional contribution exceeds the reference value.

16. The method of claim 15, wherein the reference value is determined by:

determining the second haplotype of the first chromosomal region of the first parental genome;

identifying a second set of the plurality of cell-free DNA molecules that each:

is located at any one of the one or more heterozygous loci,

includes the corresponding second allele of the heterozygous locus, and

includes at least one of the N genomic sites;

measuring N second mixture methylation levels at the N genomic sites using the second set of the plurality of cell-free DNA molecules;

measuring a second fractional contribution of the fetal tissue type in the mixture using the N second methylation levels;

determining the reference value as a sum of the second fractional contribution and a threshold value.

17. The method of claim 15, further comprising:

determining the second haplotype of the first chromosomal region of the first parental genome;

identifying a second set of the plurality of cell-free DNA molecules that each:

is located at any one of the one or more heterozygous loci,

includes the corresponding second allele of the heterozygous locus, and

includes at least one of the N genomic sites;

measuring N second mixture methylation levels at the N genomic sites using the second set of the plurality of cell-free DNA molecules;

determining a second fractional contribution of the fetal tissue type in the mixture using the N second methylation levels;

comparing the second fractional contribution to the reference value to determine whether the fetus inherited the second haplotype at the first chromosomal region, wherein the fetus inherited the second haplotype at the first chromosomal region when the second fractional contribution exceeds the reference value.

18. A method of determining a portion of a fetal genome of an unborn fetus of a pregnant female using a biological sample from the pregnant female, wherein the biological sample including a mixture of cell-free DNA molecules from a plurality of tissues types, including maternal tissue types and a fetal tissue type, the fetus having a father and a mother being the pregnant female, the method comprising:

analyzing, by a computer system, a plurality of cell-free DNA molecules from the biological sample, the plurality of cell-free DNA molecules being at least 1,000 cell-free DNA molecules, wherein analyzing a cell-free DNA molecule includes:

identifying a location of the cell-free DNA molecule in a reference human genome; and

determining a respective allele of the cell-free DNA molecule;

determining a first haplotype and a second haplotype of a first chromosomal region of a first parental genome of a first parent of the fetus;

identifying a plurality of heterozygous loci of the first chromosomal region of the first parental genome, each heterozygous locus including a corresponding first allele in the first haplotype and a corresponding second allele in the second haplotype;

identifying a first set of the plurality of cell-free DNA molecules that each:

is located at any one of the plurality of heterozygous loci,

includes the corresponding first allele of the heterozygous locus, and

includes at least one of N genomic sites;

measuring a first mixture methylation level using the first set of the plurality of cell-free DNA molecules;

identifying a second set of the plurality of cell-free DNA molecules that each:

is located at any one of the plurality of heterozygous loci,

includes the corresponding second allele of the heterozygous locus, and

includes at least one of the N genomic sites;

measuring a second mixture methylation level using the second set of the plurality of cell-free DNA molecules; and

determining which of the first haplotype and the second haplotype is inherited by the fetus based on which of the first mixture methylation level and the second mixture methylation level is lower.

19. The method of claim 18, wherein the first mixture methylation level is a methylation density of the first set of the plurality of cell-free DNA molecules, and wherein the second mixture methylation level is the methylation density of the second set of the plurality of cell-free DNA molecules.

20. The method of claim 18, wherein determining which of the first haplotype and the second haplotype is inherited includes:

determining a separation value between the first mixture methylation level and the second mixture methylation level; and

comparing the separation value to a threshold value.

21 . The method of claim 20, wherein the threshold value is determined using Student's t-test, Mann-Whitney rank-sum test, or Chi-square test.

22. The method of claim 18, wherein determining the first haplotype and the second haplotype of the first chromosomal region of the first parental genome includes:

determining genotypes of the first parent genome at the plurality of heterozygous loci using a sample from the first parent;

obtaining a plurality of reference haplotypes; and

deducing the first haplotype and the second haplotype using the genotypes and the plurality of reference haplotypes.

23. A method of detecting a sequence imbalance in a portion of a fetal genome of an unborn fetus of a pregnant female using a biological sample from the pregnant female, wherein the biological sample including a mixture of cell-free DNA molecules from a plurality of tissues types, including maternal tissue types and a fetal tissue type, the fetus having a father and a mother being the pregnant female, the method comprising:

analyzing, by a computer system, a plurality of cell-free DNA molecules from the biological sample, the plurality of cell-free DNA molecules being at least 1,000 cell-free DNA molecules, wherein analyzing a cell-free DNA molecule includes:

identifying a location of the cell-free DNA molecule in a reference human genome; and

determining a respective allele of the cell-free DNA molecule;

determining a first target haplotype of a target chromosomal region of a first parental genome of a first parent of the fetus;

identifying a plurality of target heterozygous loci of the target chromosomal region of the first parental genome, each target heterozygous locus including a corresponding first target allele in the first target haplotype and a corresponding second target allele in a second target haplotype of the target chromosomal region of the first parental genome;

identifying a target set of the plurality of cell-free DNA molecules that each:

is located at any one of the target heterozygous loci,

includes the corresponding first target allele of the target heterozygous locus, and

includes at least one of N genomic sites in the target chromosomal region, N being an integer greater than or equal to 2;

measuring N target mixture methylation levels at the N genomic sites using the target set of the plurality of cell-free DNA molecules;

determining, by the computer system, a first target fractional contribution of the fetal tissue type in the mixture using the N target methylation levels; and

determining a first reference haplotype of a reference chromosomal region of the first parental genome, the reference chromosomal region being different than the target chromosomal region;

identifying a plurality of reference heterozygous loci of the reference chromosomal region of the first parental genome, each reference heterozygous locus including a corresponding first reference allele in the first reference haplotype and a corresponding second reference allele in a second reference haplotype of the reference chromosomal region of the first parental genome;

identifying a reference set of the plurality of cell-free DNA molecules that each:

is located at any one of the reference heterozygous loci,

includes the corresponding first reference allele of the reference heterozygous locus, and

includes at least one of K genomic sites in the reference chromosomal region;

measuring K reference mixture methylation levels at the K genomic sites using the reference set of the plurality of cell-free DNA molecules;

determining, by the computer system, a first reference fractional contribution of the fetal tissue type in the mixture using the K reference methylation levels;

computing a first separation value between the first target fractional contribution and the first reference fractional contribution; and

comparing the first separation value to a threshold value to determine a classification of whether the fetus has the sequence imbalance for the target chromosomal region.

24. The method of claim 23, wherein the sequence imbalance is an amplification of the target chromosomal region.

25. The method of claim 23, wherein the threshold value used is dependent on a type of the sequence imbalance being tested.

26. The method of claim 23, further comprising:

measuring a fetal DNA percentage in the biological sample; and

using the fetal DNA percentage to determine the threshold value.

27. The method of claim 23, further comprising:

determining the threshold value using a reference group of samples, all of which have the sequence imbalance or do not have the sequence imbalance.

28. The method of claim 23, further comprising:

determining that the fetus has inherited the first target haplotype from the first parent; and

determining that the fetus has inherited the first reference haplotype from the first parent.

29. The method of claim 28, wherein determining that the fetus has inherited the first target haplotype from the first parent includes:

determining a second target fractional contribution of the fetal tissue type in the mixture corresponding to the second target haplotype;

computing a second separation value between the first target fractional contribution and the second target fractional contribution;

determining that the fetus has inherited the first target haplotype from the first parent based on the second separation value.

30. The method of claim 23, wherein the target chromosomal region is a first chromosome and the reference chromosomal region is a second chromosome different from the first chromosome.

31. The method of claim 23, wherein the first parent is the mother of the fetus.

32. A method of analyzing a biological sample of an organism, the biological sample including a mixture of cell-free DNA molecules from a plurality of tissues types, including a first tissue type, the method comprising:

analyzing, by a computer system, a plurality of cell-free DNA molecules from the biological sample, the plurality of cell-free DNA molecules being at least 1,000 cell-free DNA molecules, wherein analyzing a cell-free DNA molecule includes:

identifying a location of the cell-free DNA molecule in a reference genome corresponding to the organism;

determining a respective allele of the cell-free DNA molecule;

identifying one or more loci, each having a corresponding first allele on a first haplotype of a first chromosomal region, wherein the first haplotype is:

not present in healthy cells of the organism, or

present in the healthy cells of the organism and not present in abnormal cells that may be in the mixture;

identifying a first set of the plurality of cell-free DNA molecules that each:

is located at any one locus of the one or more loci,

includes the corresponding first allele of the one locus, and

includes at least one of N genomic sites, N being an integer greater than or equal to 2;

measuring N first mixture methylation levels at the N genomic sites using the first set of the plurality of cell-free DNA molecules;

determining, by the computer system, a first fractional contribution of the first tissue type in the mixture using the N first methylation levels;

determining a first separation value between the first fractional contribution and a reference fractional contribution; and

comparing the first separation value to a threshold value to determine a classification of whether the first tissue type has a disease state.

33. The method of claim 32, wherein the first haplotype is on a first chromosome, and wherein the first tissue type has the first haplotype on both copies of the first chromosome.

34. The method of claim 32, wherein the first haplotype is not present in the healthy cells of the organism, and wherein the first haplotype is associated with the disease state.

35. The method of claim 34, wherein the first tissue type is determined to have the disease state when the first separation value is greater than the threshold value.

36. The method of claim 34, further comprising:

identifying a second set of the plurality of cell-free DNA molecules that each:

is located at any one of the one or more loci;

includes a corresponding second allele on a second haplotype of the first chromosomal region, the second haplotype not being present in the healthy cells of the organism, and the second haplotype associated with the disease state; and

includes at least one of the N genomic sites;

measuring N second mixture methylation levels at the N genomic sites using the second set of the plurality of cell-free DNA molecules;

determining, by the computer system, a second fractional contribution of the first tissue type in the mixture using the N second methylation levels;

determining a second separation value between the second fractional contribution and the reference fractional contribution; and

comparing the second separation value to the threshold value as part of determining the classification of whether the first tissue type has the disease state.

37. The method of claim 34, wherein the disease state is cancer.

38. The method of claim 37, wherein a particular cancer has the first haplotype in a cancer genome of the particular cancer.

39. The method of claim 38, wherein the particular cancer is selected from a group consisting of liver cancer, lung cancer, pancreas cancer, atrium cancer, colon cancer, sigmoid colon cancer, transverse colon cancer, ascending colon cancer, descending colon cancer, adrenal gland cancer, esophagus cancer, small intestines cancer, and CD4 T cell cancer.

40. The method of claim 37, wherein the classification of whether the first tissue type has cancer comprises classification of the presence or absence of a cancer, classification of a stage of cancer, classification of a size of a tumor, and/or classification of metastasis.

41. The method of claim 37, further comprising performing, by the computer system:

determining a plurality of corresponding fractional contributions of other tissue types in the mixture using the N first mixture methylation levels;

determining corresponding separation values between the plurality of corresponding fractional contributions and corresponding reference fractional contributions; and

comparing the corresponding separation values to the threshold value to determine classifications of whether each of the other tissue types has the particular cancer.

42. The method of claim 34, wherein the corresponding first allele at each of the one or more loci is a cancer-specific mutation.

43. The method of claim 32, further comprising:

determining the corresponding first allele at each of the one or more loci on the first haplotype by:

analyzing DNA molecules of a tissue sample having the first haplotype to determine the first haplotype.

44. The method of claim 32, wherein the disease state is a rejection of a transplant of the first tissue type in the organism.

45. The method of claim 44, wherein the reference fractional contribution is determined from one or more measurements of biological samples of organisms with the first tissue type being healthy.

46. The method of claim 44, wherein the reference fractional contribution is determined from one or more measurements of biological samples of organisms whose transplanted first tissue is not being rejected.

47. The method of claim 32, wherein the reference fractional contribution is zero.

48. The method of claim 32, wherein the N first mixture methylation levels form a methylation vector b, and wherein determining the first fractional contribution of the first tissue type includes:

for each of M tissue types:

obtaining N tissue-specific methylation levels at the N genomic sites, N being greater than or equal to M, wherein the tissue-specific methylation levels form a matrix A of dimensions N by M, the M tissue types including the first tissue type, M being an integer greater than one;

solving for a composition vector x that provides the methylation vector b for the matrix A

for each component of one or more components of the composition vector x:

using the component to determine a corresponding fractional contribution of a corresponding tissue type of the M tissue types in the mixture.

49. The method of claim 32, wherein the first haplotype is present in the healthy cells of the organism and not present in the abnormal cells that may be in the mixture.

50. The method of claim 32, wherein the first tissue type is determined to have the disease state when the first separation value is less than the threshold value.

51. The method of claim 49, wherein the abnormal cells are from a tumor.

52. The method of claim 49, wherein the abnormal cells are from donor tissue.

53. A method of analyzing a biological sample of an organism to identify an origin of a chromosomal aberration, the biological sample including a mixture of cell-free DNA molecules from a plurality of tissues types, including a first tissue type, the method comprising:

analyzing, by a computer system, a plurality of cell-free DNA molecules from the biological sample, the plurality of cell-free DNA molecules being at least 1,000 cell-free DNA molecules, wherein analyzing a cell-free DNA molecule includes:

identifying a location of the cell-free DNA molecule in a reference genome corresponding to the organism;

determining a respective allele of the cell-free DNA molecule;

identifying a first chromosomal region as exhibiting a copy number aberration in the organism based on a first amount of cell-free DNA molecules that are located in the first chromosomal region;

determining a first haplotype and a second haplotype of the organism in the first chromosomal region;

identifying one or more heterozygous loci of the first chromosomal region, each heterozygous locus including a corresponding first allele in the first haplotype and a corresponding second allele in the second haplotype;

identifying a first set of the plurality of cell-free DNA molecules that each:

is located at any one of the one or more heterozygous loci,

includes the corresponding first allele of the heterozygous locus, and

includes at least one of N genomic sites, N being an integer greater than or equal to 2;

measuring N first mixture methylation levels at the N genomic sites using the first set of the plurality of cell-free DNA molecules;

identifying a second set of the plurality of cell-free DNA molecules that each:

is located at any one of the one or more heterozygous loci,

includes the corresponding second allele of the heterozygous locus, and

includes at least one of the N genomic sites;

measuring N second mixture methylation levels at the N genomic sites using the second set of the plurality of cell-free DNA molecules;

for each tissue type of M tissue types:

determining, by the computer system, a corresponding first fractional contribution of the tissue type in the mixture using the N first methylation levels, M being an integer greater than one;

determining, by the computer system, a corresponding second fractional contribution of the tissue type in the mixture using the N second methylation levels;

computing a corresponding separation value between the corresponding first fractional contribution and the corresponding second fractional contribution;

identifying the first tissue type as being the origin of the copy number aberration based on a first separation value of the first tissue type having a maximal value among the corresponding separate values.

54. The method of claim 53, further comprising:

determining a difference between the first separation value and a next highest separation value; and

comparing the difference to a threshold to determine a classification of how likely the first tissue type is the origin of the copy number aberration.

55. The method of claim 53, wherein identifying the first chromosomal region as exhibiting the copy number aberration includes:

comparing the first amount of cell-free DNA molecules that are located in the first chromosomal region to a reference value determined from amounts of cell-free DNA molecules that are located in other chromosomal regions.

56. The method of claim 53, wherein the copy number aberration is a copy number gain.

57. The method of claim 53, further comprising:

determining a first number of cell-free DNA molecules in the first set of the plurality of cell-free DNA molecules;

determining a second number of cell-free DNA molecules in the second set of the plurality of cell-free DNA molecules; and

using the first number and the second number to determine whether the maximal value is to be positive or negative.

58. The method of claim 53, wherein analyzing the cell-free DNA molecule includes: determining a size of the cell-free DNA molecule, wherein the first set of the plurality of cell-free DNA molecules has a first size distribution, and wherein the second set of the plurality of cell-free DNA molecules has a second size distribution, the method further comprising:

determining a first statistical value of the first size distribution;

determining a second statistical value of the second size distribution; and

using the first statistical value and the second statistical value to determine whether the maximal value is to be positive or negative.

59. The method of claim 58, further comprising:

determining which of the first haplotype and the second haplotype has a higher copy number using the first statistical value and the second statistical value; and

computing the corresponding separation values such that the first separation value is positive.

60. The method of any one of claims 1-59, wherein analyzing the cell-free DNA molecules comprises bisulfite sequencing, sequencing preceded by methylation-sensitive restriction enzyme digestion, immunoprecipitation using anti-methylcytosine antibody or methylation binding protein, or single molecule sequencing that allows elucidation of the methylation status.

61. A computer product comprising a computer readable medium storing a plurality of instructions for controlling a computer system to perform an operation of any method of claims 1-59.

62. A system comprising:

the computer product of claim 61; and

one or more processors for executing instructions stored on the computer readable medium.

63. A system comprising means for performing any method of claims 1-59 .

64. A system configured to perform any method of claims 1-59.

65. A system comprising modules that respectively perform the steps of any method of claims 1-59.

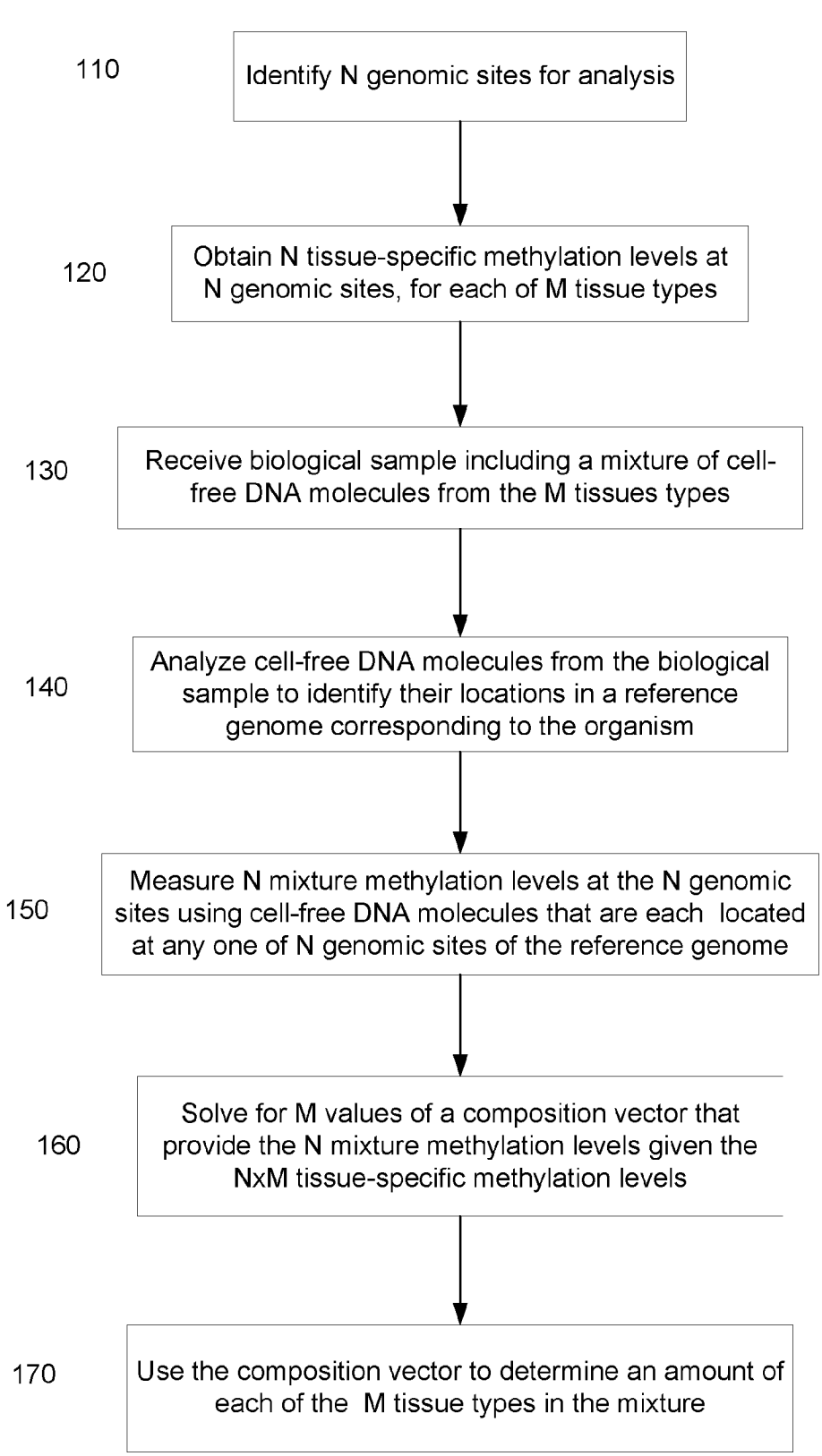

100
110
Identify N genomic sites for analysis
120
Obtain N tissue-specific methylation levels at N genomic sites, for each of M tissue types
130
Receive biological sample including a mixture of cell-free DNA molecules from the M tissues types
140
Analyze cell-free DNA molecules from the biological sample to identify their locations in a reference genome corresponding to the organism
150
Measure N mixture methylation levels at the N genomic sites using cell-free DNA molecules that are each located at any one of N genomic sites of the reference genome
160
Solve for M values of a composition vector that provide the N mixture methylation levels given the NxM tissue-specific methylation levels
170
Use the composition vector to determine an amount of each of the M tissue types in the mixture

FIG. 1

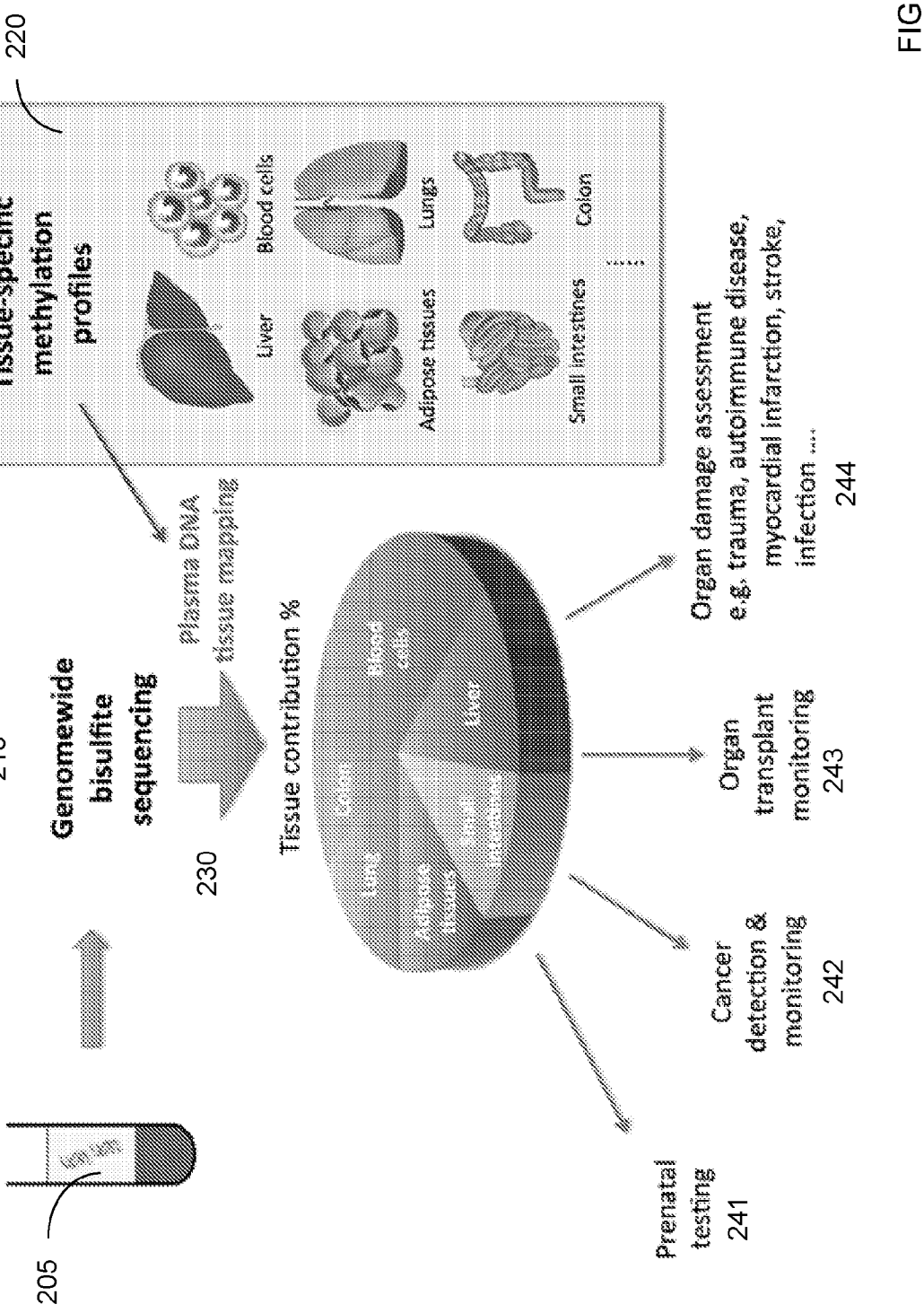

205
210
Genomewide bisulfite sequencing
230
Plasma DNA tissue mapping
Tissue contribution %
220
Tissue-specific methylation profiles
Liver
Blood cells
Adipose tissues
Lungs
Small intestines
Colon
Prenatal testing 241
Cancer detection & monitoring 242
Organ transplant monitoring 243
Organ damage assessment e.g. trauma, autoimmune disease, myocardial infarction, stroke, infection ... 244

FIG. 2

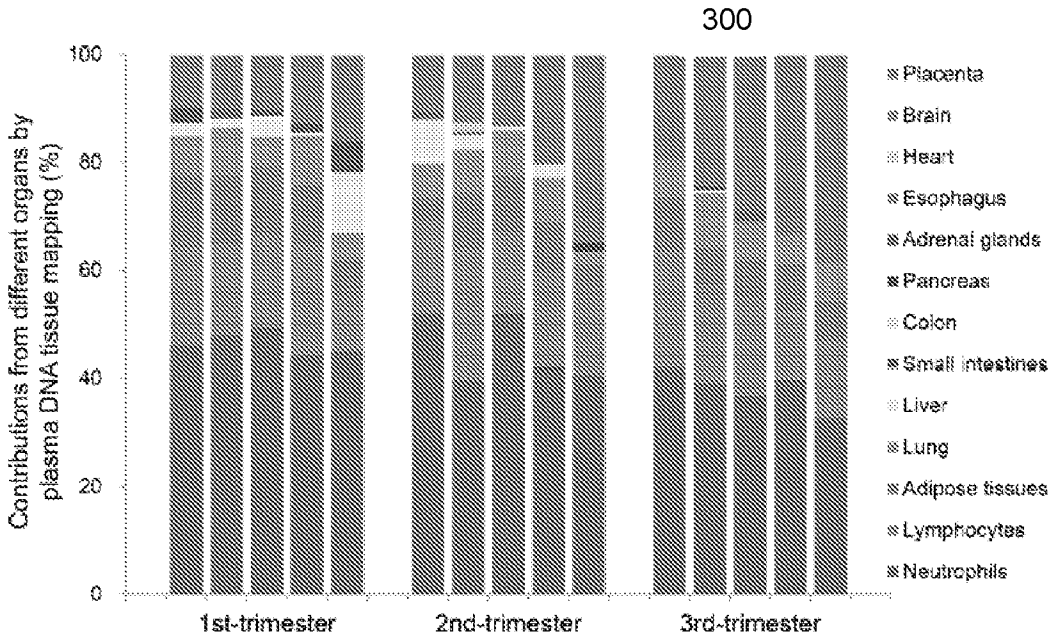

300
Contributions from different organs by plasma DNA tissue mapping (%)
100
80
60
40
20
0
1st-trimester
2nd-trimester
3rd-trimester
Placenta
Brain
Heart
Esophagus
Adrenal glands
Pancreas
Colon
Small intestines
Liver
Lung
Adipose tissues
Lymphocytes
Neutrophils

FIG. 3A

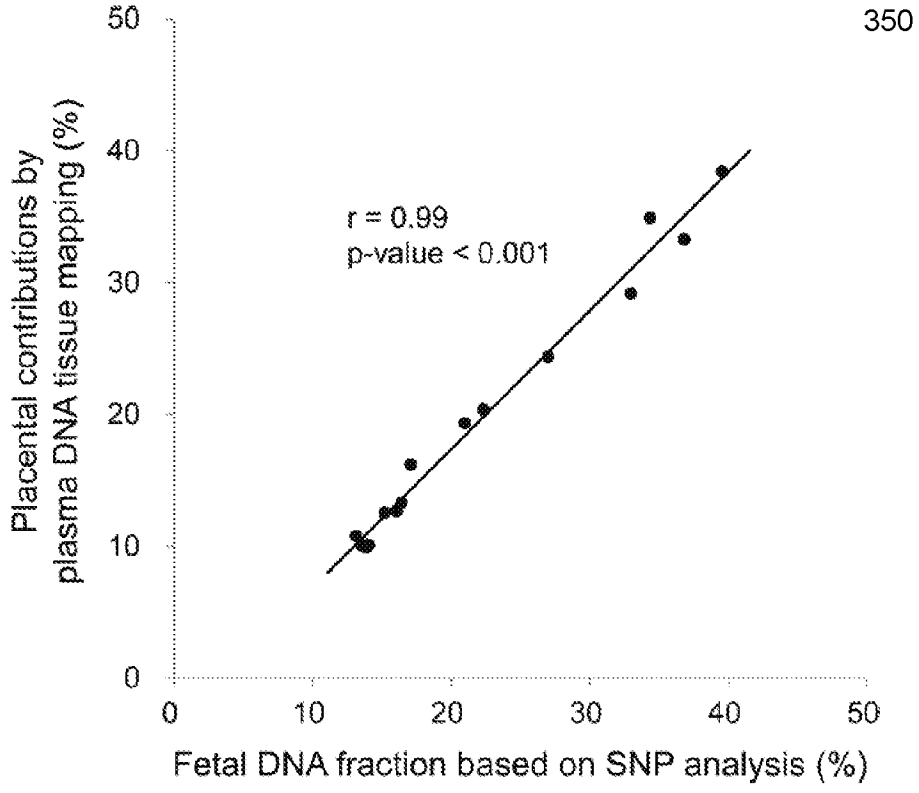

350
Placental contributions by plasma DNA tissue mapping (%)
50
40
30
20
10
0
r = 0.99
p-value < 0.001
0
10
20
30
40
50
Fetal DNA fraction based on SNP analysis (%)

FIG. 3B

400

Case no.	Percentage contributions of plasma DNA (%)													SNP-based fetal DNA fraction (%)	Gestational age (weeks)
	Liver	Lungs	Colon	Small intestines	Pancreas	Adrenal glands	Esophagus	Adipose tissues	Heart	Brain	Neutrophils	Lymphocytes	Placenta		
1	1.5	9.3	0.0	0.0	0.0	1.9	0.0	12.1	0.0	0.0	48.1	17.0	10.0	13.6	13 1/7
2	2.5	7.6	0.0	2.8	0.0	0.0	0.0	12.6	0.0	0.0	46.1	18.6	9.9	13.9	13
3	3.8	5.8	0.0	0.7	0.0	0.6	0.0	12.4	0.0	0.0	49.2	17.4	10.1	14.0	12 6/7
4	0.6	9.2	0.0	1.8	0.0	0.0	0.0	10.6	0.0	0.0	44.1	21.1	12.7	16.1	13
5	11.1	5.2	0.0	5.7	0.0	0.0	0.0	5.1	0.0	0.0	45.0	11.7	16.2	17.1	12 6/7
6	8.1	6.4	0.0	0.0	0.0	1.3	0.0	4.5	0.0	0.0	51.9	17.0	10.8	13.1	21 3/7
7	3.0	8.2	0.0	0.0	0.0	0.5	0.0	10.1	1.7	0.0	39.5	24.5	12.5	15.2	20 6/7
8	0.8	8.5	0.0	0.0	0.0	0.0	0.0	10.2	0.0	0.0	51.7	15.5	13.3	16.4	22 2/7
9	2.6	8.0	0.0	0.0	0.0	0.0	0.0	9.2	0.0	0.0	41.7	18.1	20.4	22.3	22 1/7
10	0.0	2.6	0.0	0.0	1.6	0.0	0.0	4.1	0.0	0.0	40.5	16.3	34.9	34.3	21 /7
11	0.0	7.4	0.0	0.0	0.0	0.0	0.0	9.8	0.0	0.0	41.5	22.0	19.3	21.0	38 1/7
12	0.6	10.1	0.0	0.8	0.0	0.0	0.0	6.6	0.0	0.0	39.0	18.7	24.3	27.0	38 3/7
13	0.0	7.4	0.0	0.0	0.0	1.7	0.0	6.3	0.0	0.0	36.7	18.7	29.2	33.0	38 1/7
14	0.0	5.9	0.0	0.0	0.0	0.0	0.0	8.3	0.0	0.0	39.6	12.9	33.3	36.8	38 2/7
15	0.0	7.3	0.0	0.0	0.0	0.0	0.0	7.2	0.0	0.0	32.5	14.6	38.4	39.5	38 3/7

FIG. 4

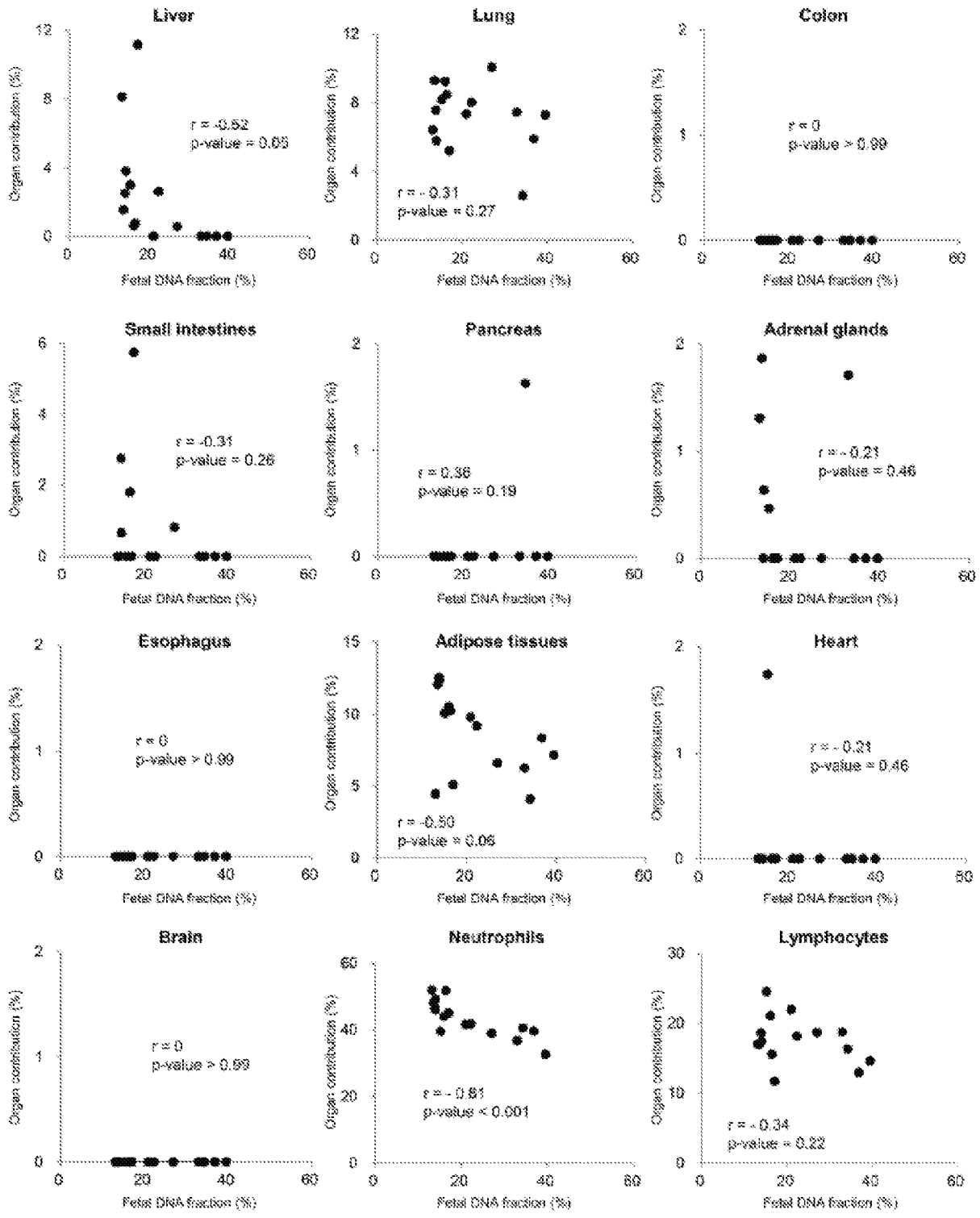

Liver
r = -0.52
p-value = 0.05
Lung
r = - 0.31
p-value = 0.27
Colon
r = 0
p-value > 0.99
Small intestines
r = -0.31
p-value = 0.26
Pancreas
r = 0.36
p-value = 0.19
Adrenal glands
r = - 0.21
p-value = 0.46
Esophagus
r = 0
p-value > 0.99
Adipose tissues
r = -0.50
p-value = 0.06
Heart
r = - 0.21
p-value = 0.46
Brain
r = 0
p-value > 0.99
Neutrophils
r = - 0.81
p-value < 0.001
Lymphocytes
r = - 0.34
p-value = 0.22
Organ contribution (%)
Fetal DNA fraction (%)

FIG. 5

600

Case no.	Percentage contributions of plasma DNA (%)												
	Liver	Lungs	Colon	Small intestines	Pancreas	Adrenal glands	Esophagus	Adipose tissues	Heart	Brain	Neutrophils	Lymphocytes	Placenta
1	5.8	6.8	0.0	1.3	0.0	0.0	0.0	6.2	2.9	0.0	58.2	18.8	0.0
2	5.1	5.2	0.0	7.3	0.0	0.0	0.0	4.5	0.0	0.0	61.8	16.1	0.0
3	5.6	7.8	0.0	2.1	0.0	0.0	0.0	6.5	0.6	0.0	58.8	18.6	0.0
4	4.1	6.4	0.0	4.5	0.0	0.0	0.0	5.9	1.3	0.4	59.0	18.4	0.0
5	4.4	2.5	0.7	6.4	0.0	0.0	0.0	5.5	2.7	0.6	59.6	17.1	0.3
6	5.8	4.1	0.0	5.3	0.0	0.0	0.0	5.2	0.5	0.8	55.2	22.7	0.5
7	4.9	4.1	0.0	4.5	0.0	0.0	0.0	6.3	0.0	0.0	60.2	20.0	0.0
8	4.4	7.4	0.0	5.3	0.0	0.0	0.0	3.1	0.3	0.3	61.7	17.5	0.0
9	4.5	9.4	0.0	4.0	0.0	0.0	0.0	6.2	0.1	1.2	55.9	18.6	0.0
10	4.6	0.1	0.2	1.3	0.0	0.0	0.0	7.5	0.0	2.4	60.0	23.9	0.0
11	4.7	4.5	0.0	5.5	0.0	0.0	0.0	5.6	0.9	0.5	62.4	16.0	0.0
12	7.3	4.9	0.0	4.4	0.0	0.0	0.0	6.0	1.8	0.9	54.0	20.7	0.0
13	2.4	10.9	0.0	1.7	0.0	0.0	0.0	5.7	0.0	0.6	53.9	24.5	0.3
14	3.8	3.5	0.0	0.0	0.0	0.0	0.0	8.9	0.0	0.0	61.2	22.7	0.0
15	5.3	7.2	0.0	4.2	0.0	0.2	0.0	5.1	0.0	0.1	56.3	21.6	0.0
16	4.9	6.5	0.0	7.9	0.0	0.0	0.0	4.3	0.5	1.2	55.4	18.9	0.4
17	18.6	12.8	0.0	0.0	0.0	0.0	0.0	4.4	0.0	0.0	47.7	16.5	0.0
18	8.3	5.0	0.0	2.1	0.0	0.0	0.0	7.5	0.0	0.2	56.2	19.4	1.3
19	6.0	0.0	1.0	2.3	0.0	0.0	0.0	9.3	0.0	0.0	54.5	26.3	0.6
20	2.0	4.8	0.0	0.0	0.0	0.5	0.0	10.5	0.0	0.0	58.5	22.5	1.1
21	8.5	0.0	0.0	4.5	0.0	0.0	0.0	9.4	0.0	0.0	53.9	23.4	0.2
22	6.4	11.1	0.0	4.0	0.0	0.0	0.0	6.7	0.0	0.0	53.7	18.1	0.0
23	2.0	2.8	0.0	8.7	0.0	0.0	0.0	6.5	1.1	0.0	62.5	15.0	1.3
24	5.4	2.2	0.0	2.7	0.0	0.0	0.0	6.9	0.0	1.6	65.2	15.2	0.9
25	21.0	0.0	0.0	7.1	0.0	0.0	0.0	3.8	0.0	0.0	51.6	16.4	0.0
26	7.0	6.4	0.0	6.1	0.0	0.0	0.0	4.1	1.6	0.0	55.9	19.0	0.0
27	7.9	10.4	0.0	0.3	0.0	0.0	0.0	4.8	0.0	0.0	59.1	17.6	0.0
28	5.4	9.7	0.0	4.2	0.0	0.0	0.0	5.1	0.0	0.0	55.7	20.0	0.0
29	6.2	6.6	0.0	0.0	0.0	0.0	0.0	6.3	0.0	0.0	60.7	20.2	0.0
30	6.9	5.1	0.0	5.2	0.0	0.0	0.0	6.4	1.4	0.0	58.8	16.2	0.0
31	14.0	6.4	0.0	4.5	0.0	0.0	0.0	4.9	0.7	0.0	53.2	16.2	0.0
32	9.9	5.5	0.0	1.9	0.0	0.0	0.0	6.3	0.6	0.0	60.1	15.6	0.0
Median	5.5	5.3	0.0	4.2	0.0	0.0	0.0	6.1	0.0	0.0	58.3	18.7	0.0
Lower quantile	4.6	3.9	0.0	1.9	0.0	0.0	0.0	5.0	0.0	0.0	55.0	16.5	0.0
Upper quantile	7.1	7.3	0.0	5.3	0.0	0.0	0.0	6.6	0.8	0.5	60.1	21.0	0.3

FIG. 6

700

	Pregnant women											Non-pregnant subjects			
	1	2	3	4	5	6	7	8	9	10	11	1	2	3	4
Buffy coat	58%	64%	65%	53%	83%	52%	75%	71%	52%	66%	69%	87%	83%	96%	80%
Esophagus	0%	0%	0%	2%	0%	0%	0%	3%	0%	0%	0%	0%	0%	0%	0%
Small intestines	0%	0%	0%	4%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Colon	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	5%	0%	0%
Pancreas	0%	1%	0%	1%	0%	0%	0%	0%	1%	3%	2%	1%	0%	0%	0%
Liver	3%	8%	10%	1%	0%	11%	6%	7%	4%	9%	8%	7%	7%	3%	6%
Lung	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Heart	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	10%	0%	0%	0%	2%
Adrenal gland	0%	1%	0%	1%	0%	0%	0%	0%	0%	0%	1%	0%	0%	0%	5%
Hippocampus	1%	3%	0%	1%	2%	4%	1%	4%	3%	3%	4%	2%	5%	0%	5%
Placenta	38%	24%	25%	36%	15%	34%	18%	15%	40%	19%	7%	2%	0%	1%	2%
Fetal DNA fraction determined by counting fetal specific alleles	32%	21%	26%	28%	11%	25%	17%	13%	13%	32%	3%	0%	0%	0%	0%

FIG. 7

800

	Pregnant women											Non-pregnant subjects			
	1	2	3	4	5	6	7	8	9	10	11	1	2	3	4
Buffy coat	43%	44%	47%	40%	47%	41%	38%	41%	43%	39%	45%	10%	11%	1%	13%
Esophagus	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Small intestines	0%	4%	1%	2%	3%	0%	1%	5%	0%	0%	0%	0%	0%	0%	0%
Colon	25%	18%	10%	23%	14%	23%	24%	20%	27%	25%	16%	17%	7%	22%	27%
Pancreas	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Liver	4%	8%	11%	4%	8%	11%	9%	0%	0%	0%	0%	0%	0%	0%	0%
Lung	0%	0%	0%	1%	0%	0%	0%	3%	0%	0%	11%	0%	0%	0%	0%
Heart	0%	0%	1%	3%	0%	0%	1%	0%	0%	0%	0%	0%	0%	0%	0%
Adrenal gland	0%	3%	6%	0%	0%	0%	7%	4%	3%	7%	7%	24%	23%	20%	4%
Hippocampus	7%	10%	7%	11%	12%	10%	8%	15%	9%	11%	16%	44%	42%	41%	50%
Placenta	21%	13%	18%	16%	15%	16%	11%	12%	18%	18%	6%	4%	16%	15%	7%
Fetal DNA fraction determined by counting fetal specific alleles	32%	21%	26%	28%	11%	25%	17%	13%	13%	32%	3%	0%	0%	0%	0%

FIG. 8

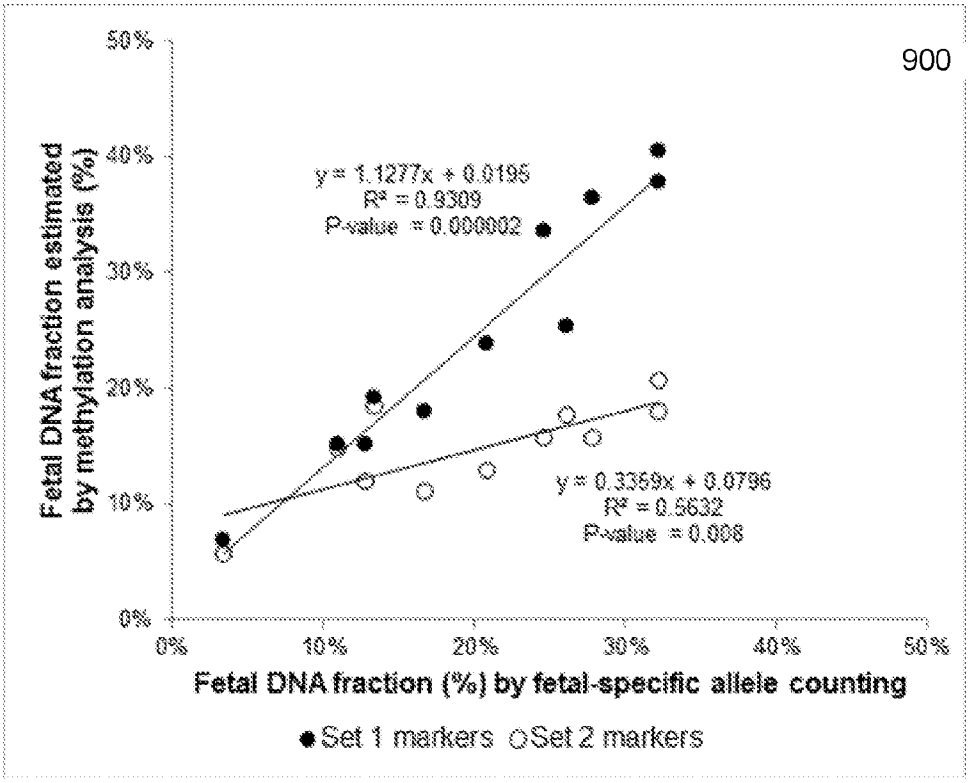

900
50%
40%
30%
20%
10%
0%
Fetal DNA fraction estimated by methylation analysis (%)
y = 1.1277x + 0.0195
R² = 0.9309
P-value = 0.000002
y = 0.3359x + 0.0796
R² = 0.5632
P-value = 0.008
0%
10%
20%
30%
40%
50%
Fetal DNA fraction (%) by fetal-specific allele counting
Set 1 markers
Set 2 markers

FIG. 9A

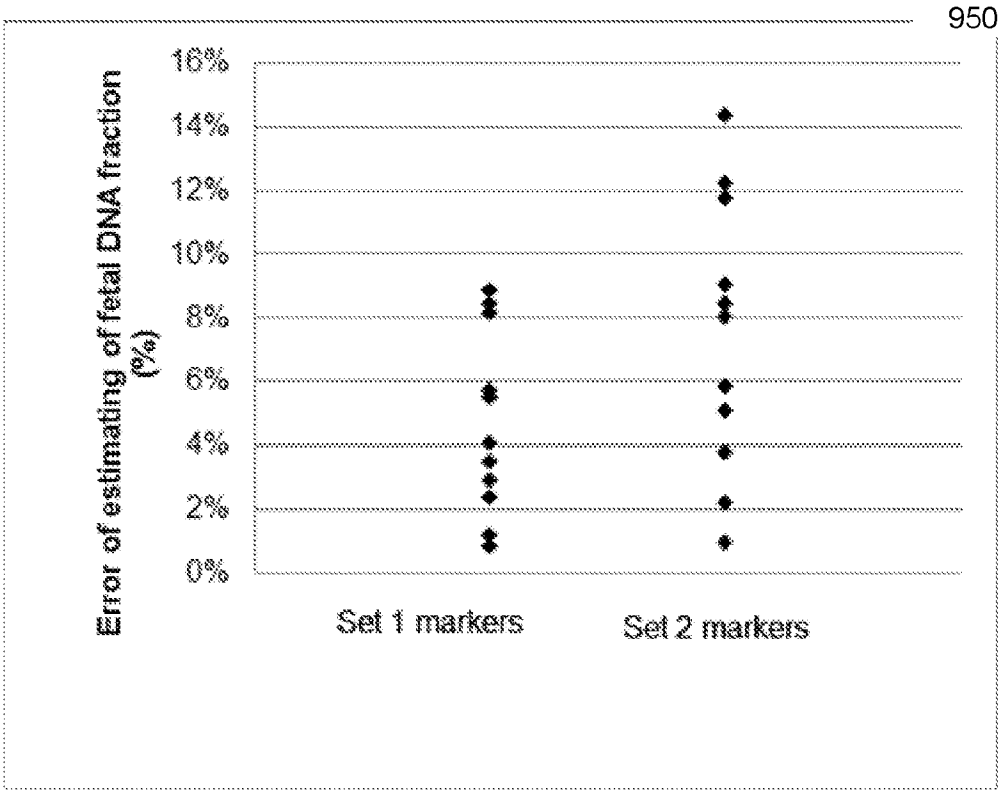

950
16%
14%
12%
10%
8%
6%
4%
2%
0%
Error of estimating of fetal DNA fraction (%)
Set 1 markers
Set 2 markers

FIG. 9B

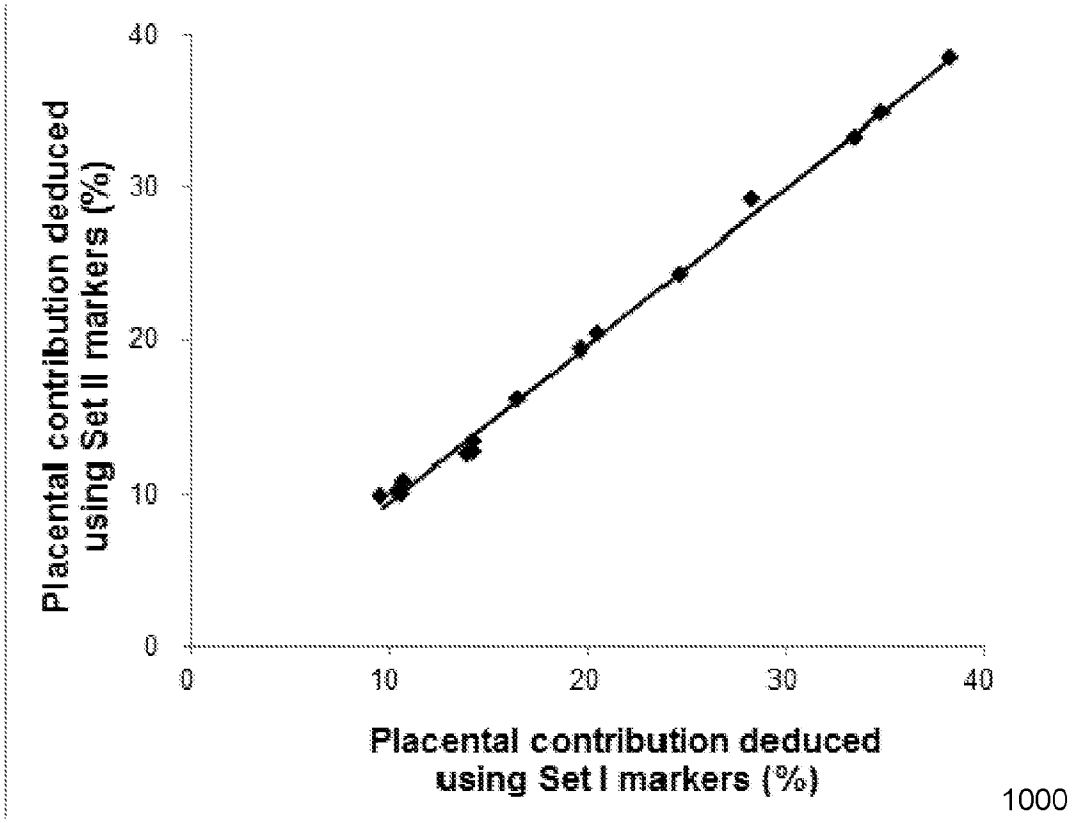


FIG. 10A

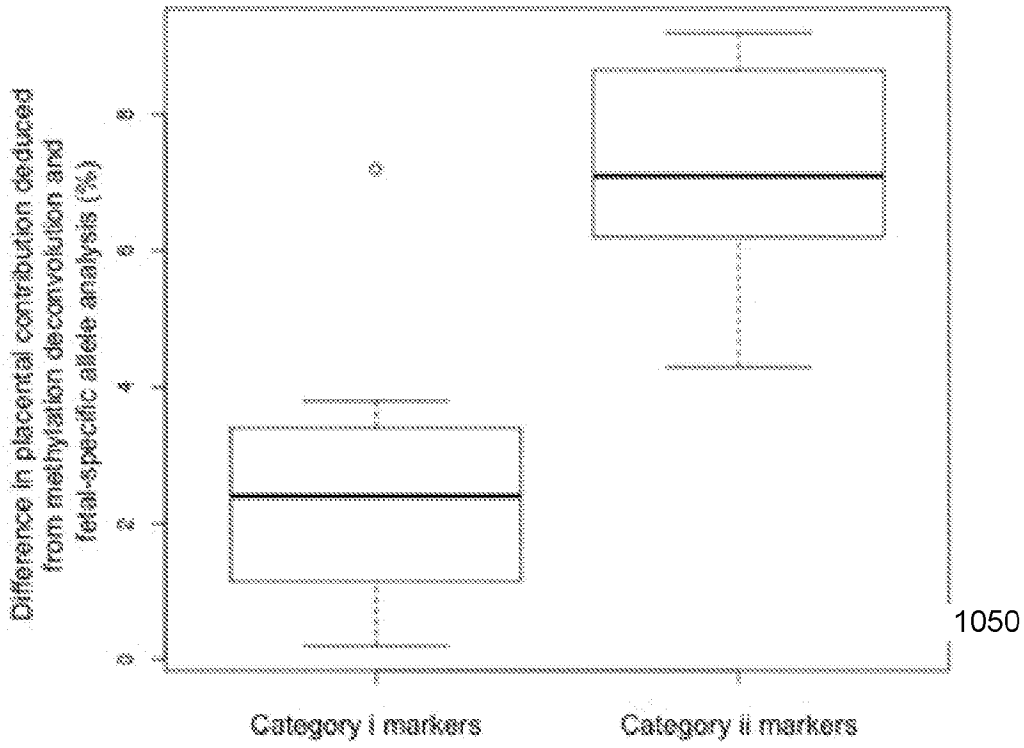


FIG. 10B

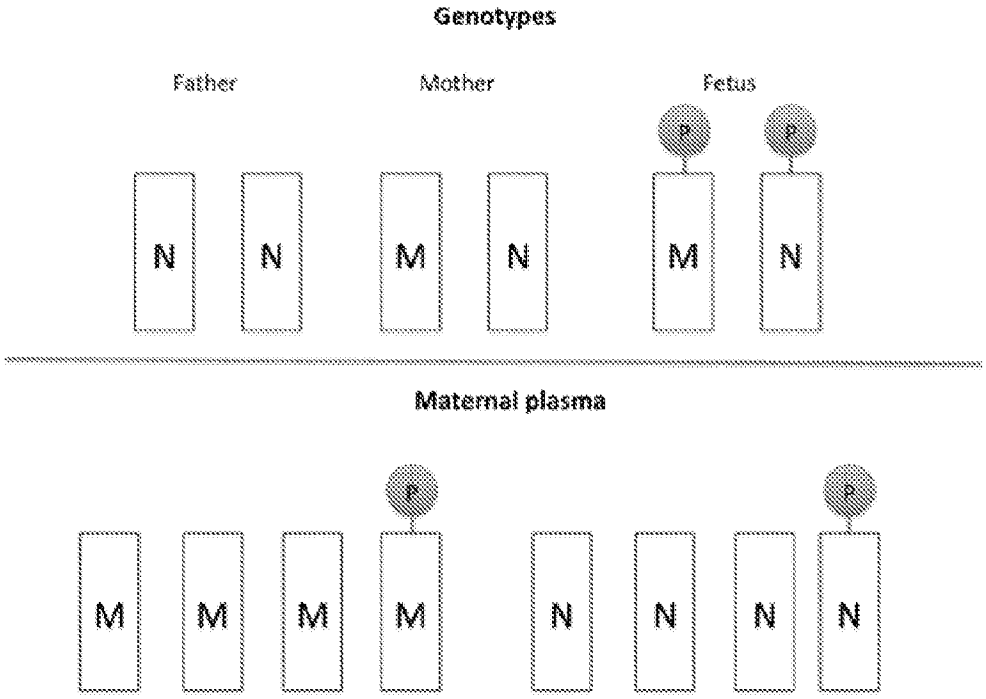


FIG. 11A

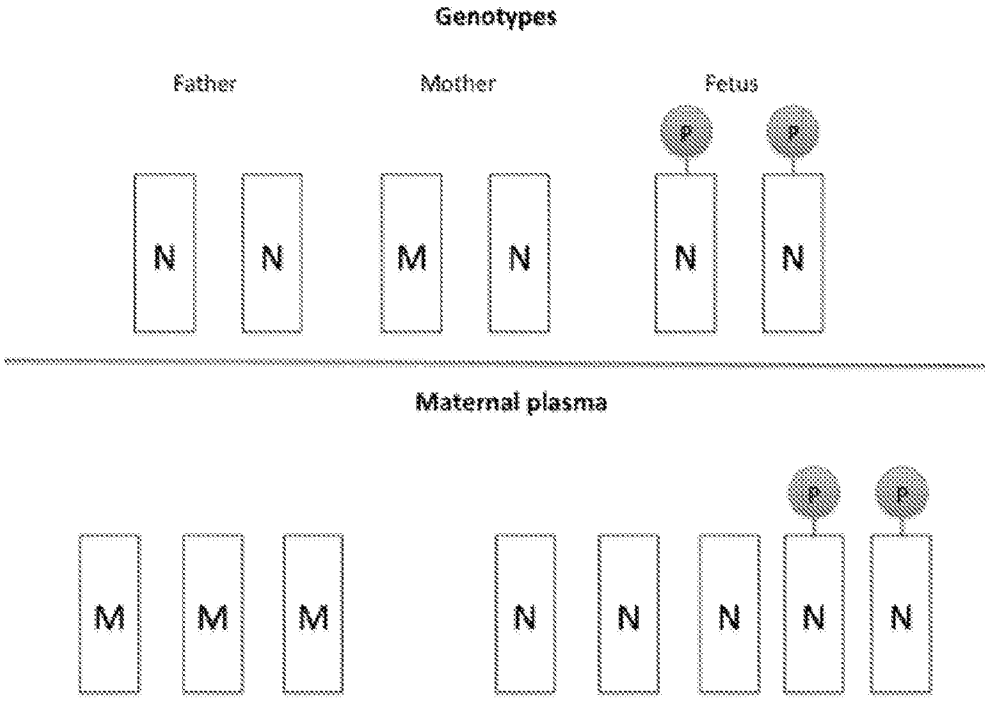


FIG. 11B

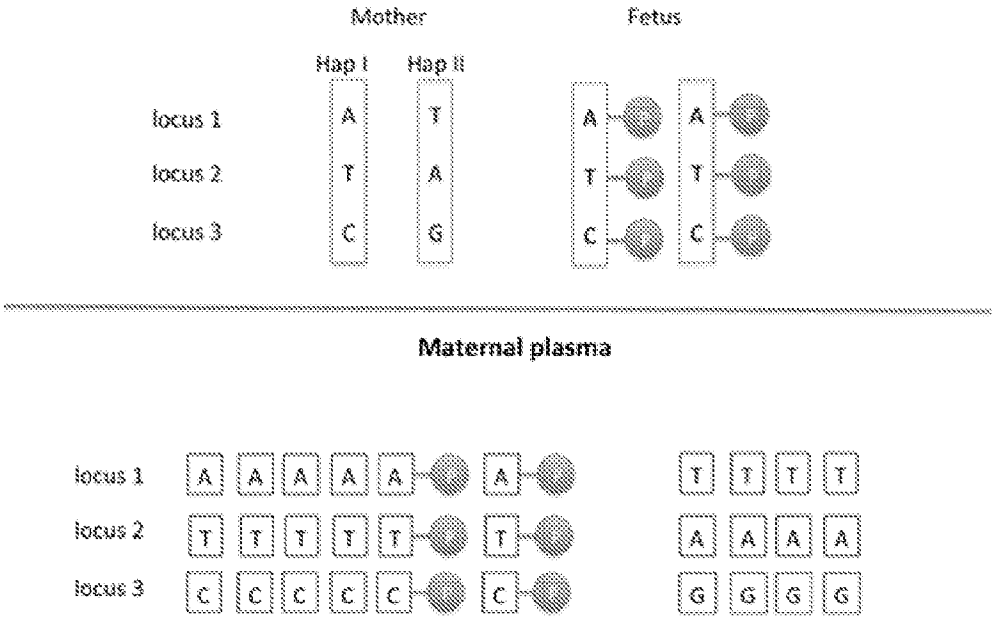


FIG. 12A

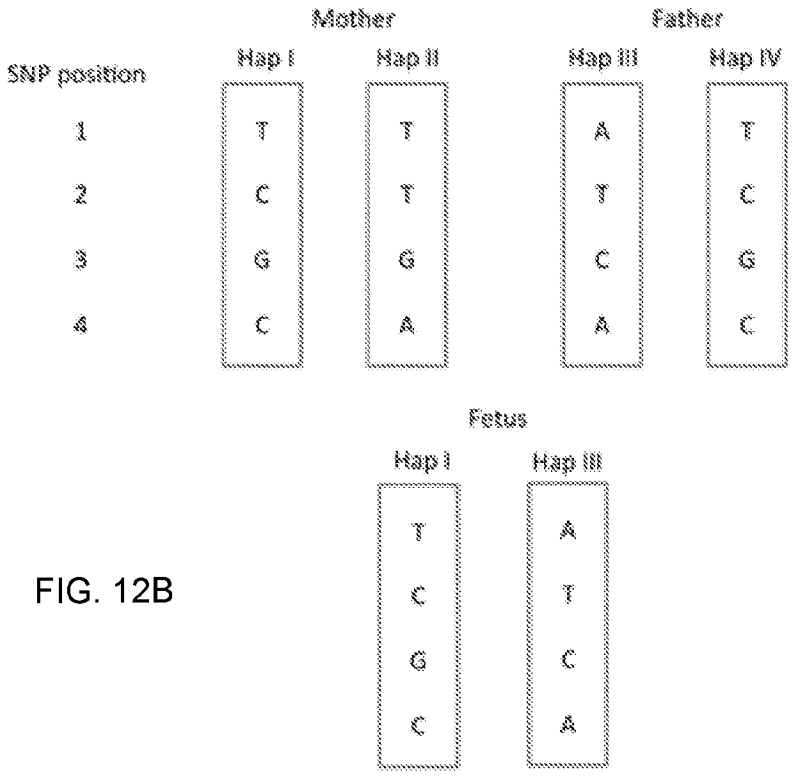


FIG. 12B

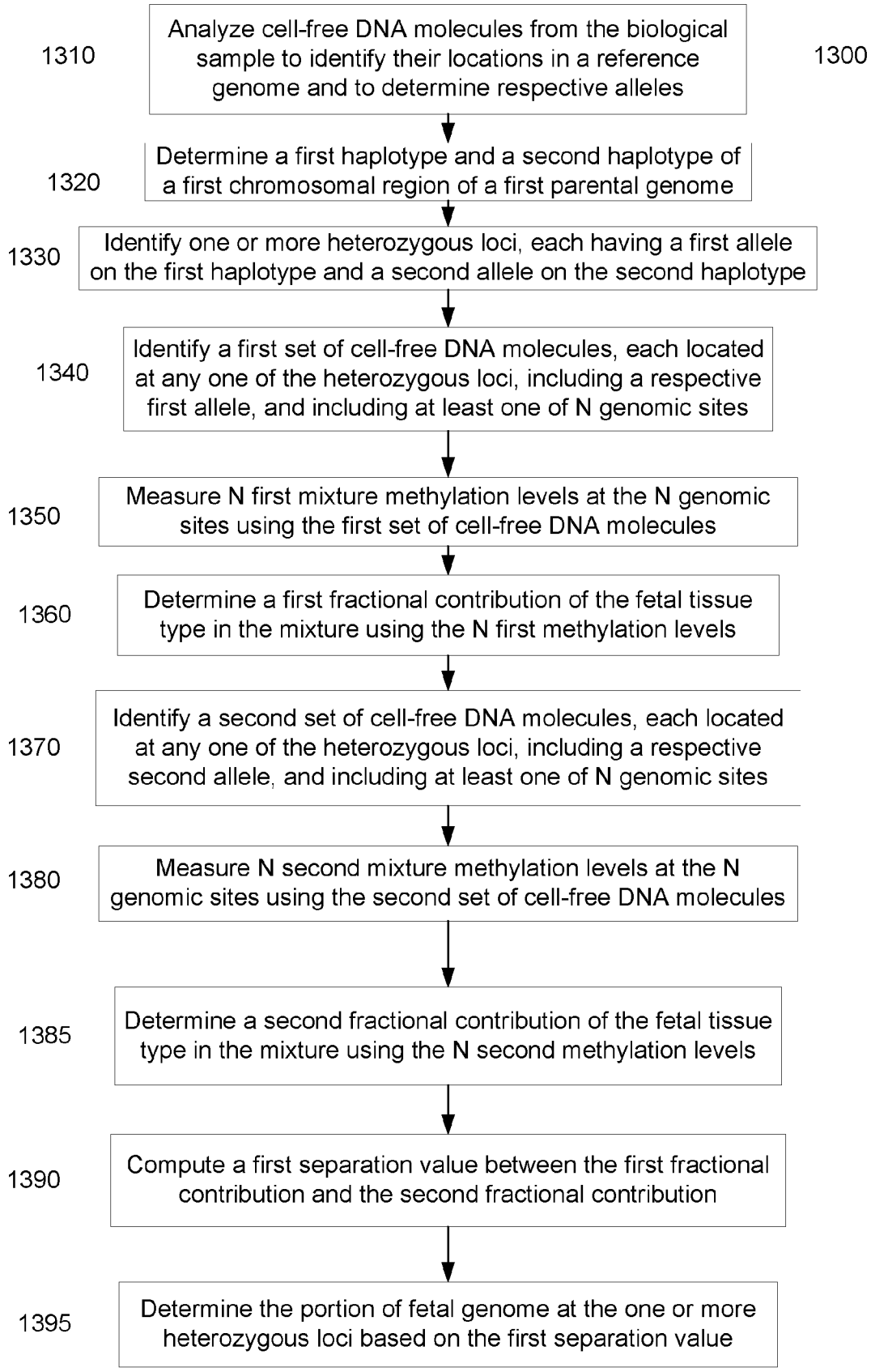

1300
1310
Analyze cell-free DNA molecules from the biological sample to identify their locations in a reference genome and to determine respective alleles
1320
Determine a first haplotype and a second haplotype of a first chromosomal region of a first parental genome
1330
Identify one or more heterozygous loci, each having a first allele on the first haplotype and a second allele on the second haplotype
1340
Identify a first set of cell-free DNA molecules, each located at any one of the heterozygous loci, including a respective first allele, and including at least one of N genomic sites
1350
Measure N first mixture methylation levels at the N genomic sites using the first set of cell-free DNA molecules
1360
Determine a first fractional contribution of the fetal tissue type in the mixture using the N first methylation levels
1370
Identify a second set of cell-free DNA molecules, each located at any one of the heterozygous loci, including a respective second allele, and including at least one of N genomic sites
1380
Measure N second mixture methylation levels at the N genomic sites using the second set of cell-free DNA molecules
1385
Determine a second fractional contribution of the fetal tissue type in the mixture using the N second methylation levels
1390
Compute a first separation value between the first fractional contribution and the second fractional contribution
1395
Determine the portion of fetal genome at the one or more heterozygous loci based on the first separation value

FIG. 13

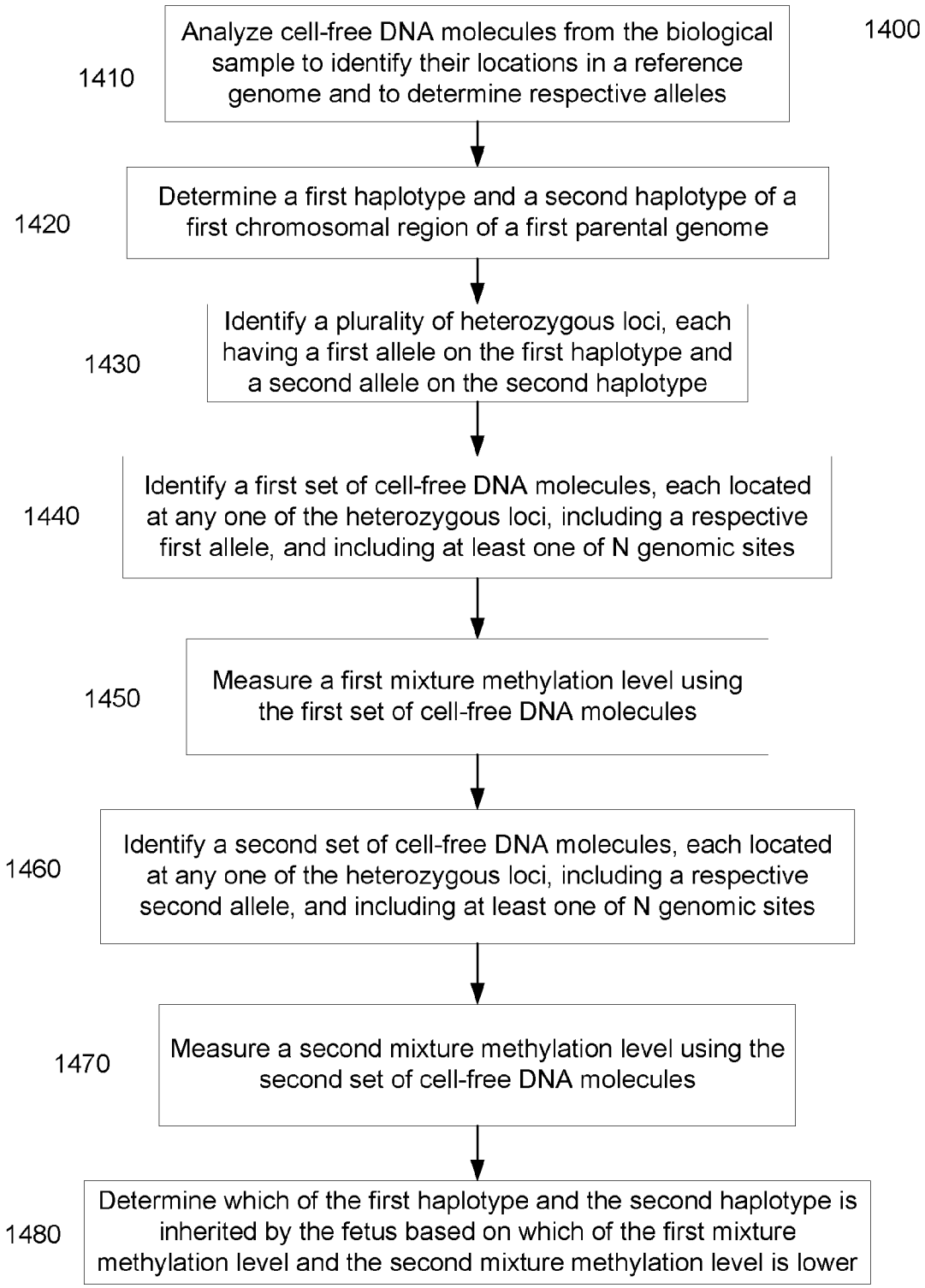

1400
1410
Analyze cell-free DNA molecules from the biological sample to identify their locations in a reference genome and to determine respective alleles
1420
Determine a first haplotype and a second haplotype of a first chromosomal region of a first parental genome
1430
Identify a plurality of heterozygous loci, each having a first allele on the first haplotype and a second allele on the second haplotype
1440
Identify a first set of cell-free DNA molecules, each located at any one of the heterozygous loci, including a respective first allele, and including at least one of N genomic sites
1450
Measure a first mixture methylation level using the first set of cell-free DNA molecules
1460
Identify a second set of cell-free DNA molecules, each located at any one of the heterozygous loci, including a respective second allele, and including at least one of N genomic sites
1470
Measure a second mixture methylation level using the second set of cell-free DNA molecules
1480
Determine which of the first haplotype and the second haplotype is inherited by the fetus based on which of the first mixture methylation level and the second mixture methylation level is lower

FIG. 14

FIG. 15

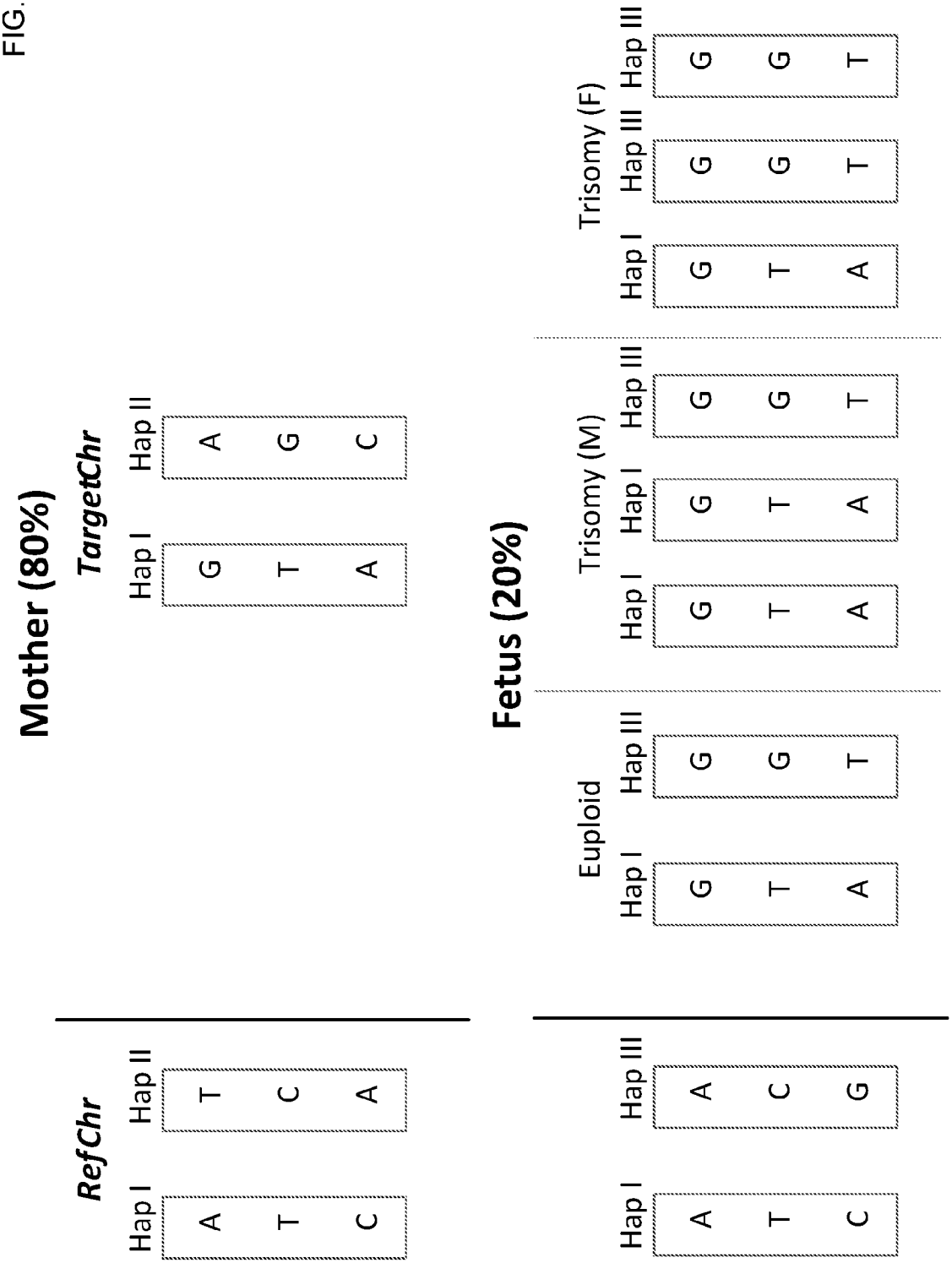
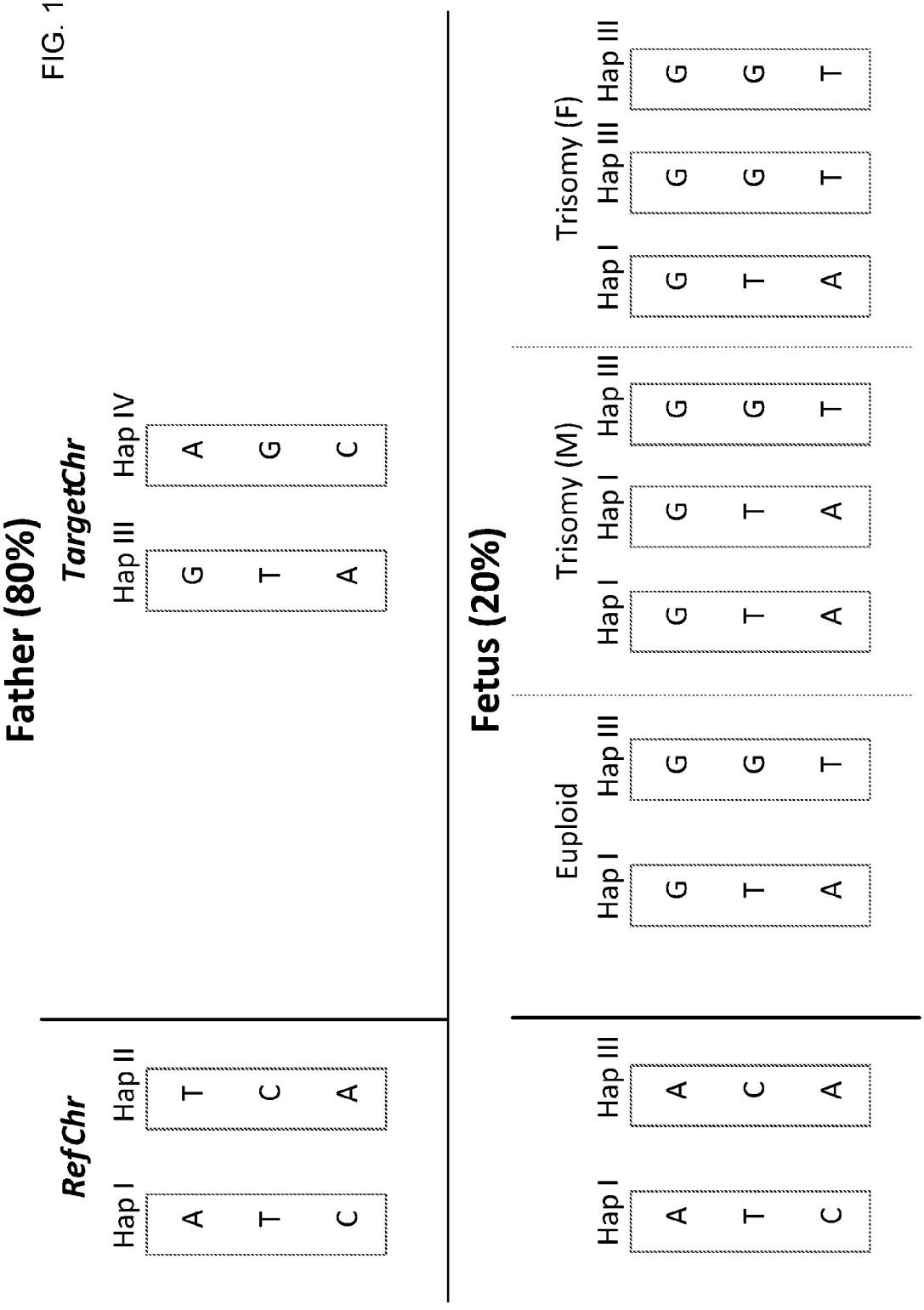

Mother (80%)
TargetChr
Hap I
Hap II
G
T
A
A
G
C
RefChr
Hap I
Hap II
A
T
C
T
C
A
Fetus (20%)
Euploid
Hap I
Hap III
G
T
A
G
G
T
Trisomy (M)
Hap I
Hap I
Hap III
G
T
A
G
T
A
G
G
T
Trisomy (F)
Hap I
Hap III
Hap III
G
T
A
G
G
T
G
G
T
Hap I
Hap III
A
T
C
A
C
G

FIG. 16


Father (80%)
RefChr
Hap I
A
T
C
Hap II
T
C
A
TargetChr
Hap III
G
T
A
Hap IV
A
G
C
Fetus (20%)
Hap I
A
T
C
Hap III
A
C
A
Euploid
Hap I
G
T
A
Hap III
G
G
T
Trisomy (M)
Hap I
G
T
A
Hap III
G
G
T
Trisomy (F)
Hap I
G
T
A
Hap III
G
G
T

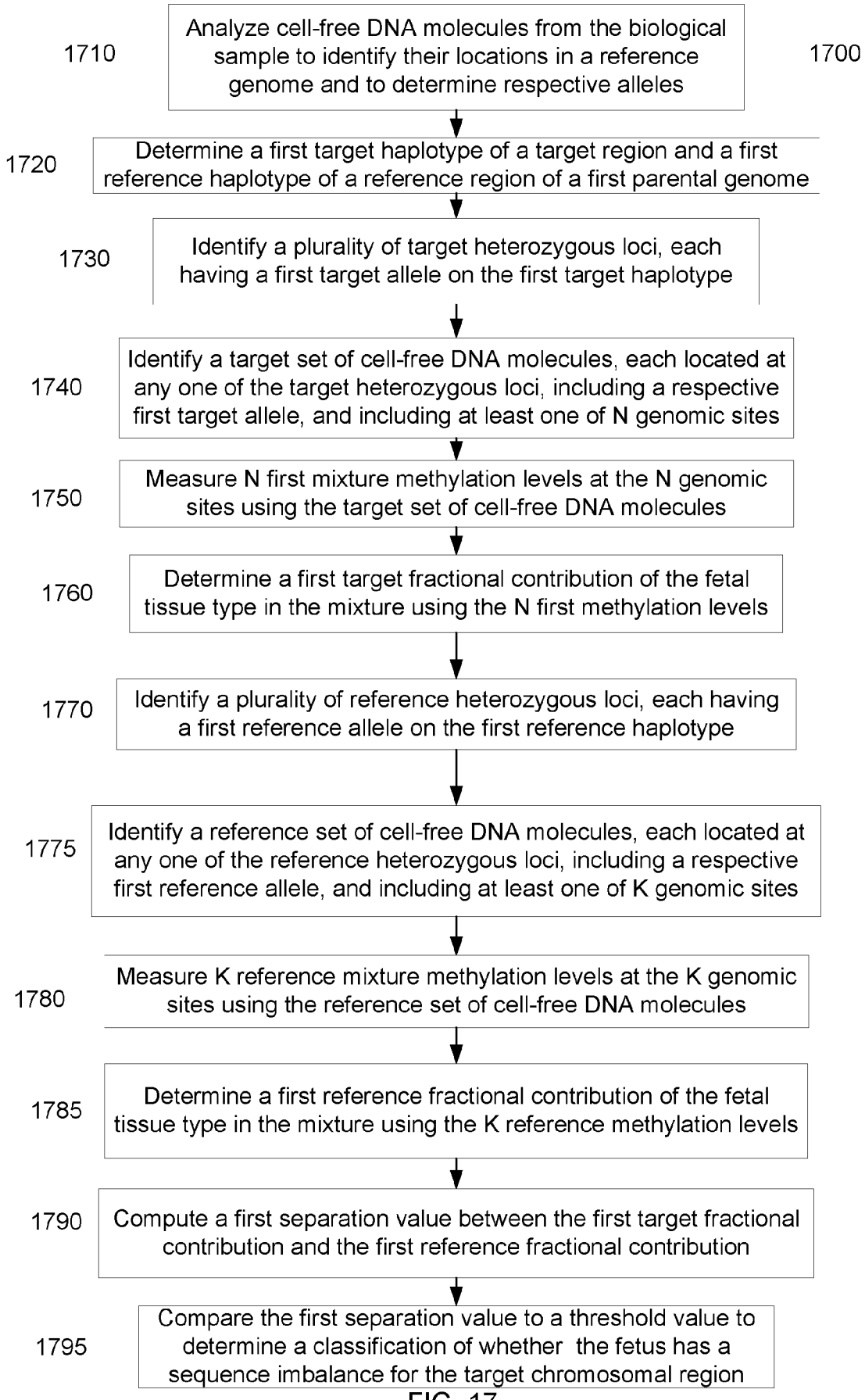

1700
1710
Analyze cell-free DNA molecules from the biological sample to identify their locations in a reference genome and to determine respective alleles
1720
Determine a first target haplotype of a target region and a first reference haplotype of a reference region of a first parental genome
1730
Identify a plurality of target heterozygous loci, each having a first target allele on the first target haplotype
1740
Identify a target set of cell-free DNA molecules, each located at any one of the target heterozygous loci, including a respective first target allele, and including at least one of N genomic sites
1750
Measure N first mixture methylation levels at the N genomic sites using the target set of cell-free DNA molecules
1760
Determine a first target fractional contribution of the fetal tissue type in the mixture using the N first methylation levels
1770
Identify a plurality of reference heterozygous loci, each having a first reference allele on the first reference haplotype
1775
Identify a reference set of cell-free DNA molecules, each located at any one of the reference heterozygous loci, including a respective first reference allele, and including at least one of K genomic sites
1780
Measure K reference mixture methylation levels at the K genomic sites using the reference set of cell-free DNA molecules
1785
Determine a first reference fractional contribution of the fetal tissue type in the mixture using the K reference methylation levels
1790
Compute a first separation value between the first target fractional contribution and the first reference fractional contribution
1795
Compare the first separation value to a threshold value to determine a classification of whether the fetus has a sequence imbalance for the target chromosomal region

FIG. 17

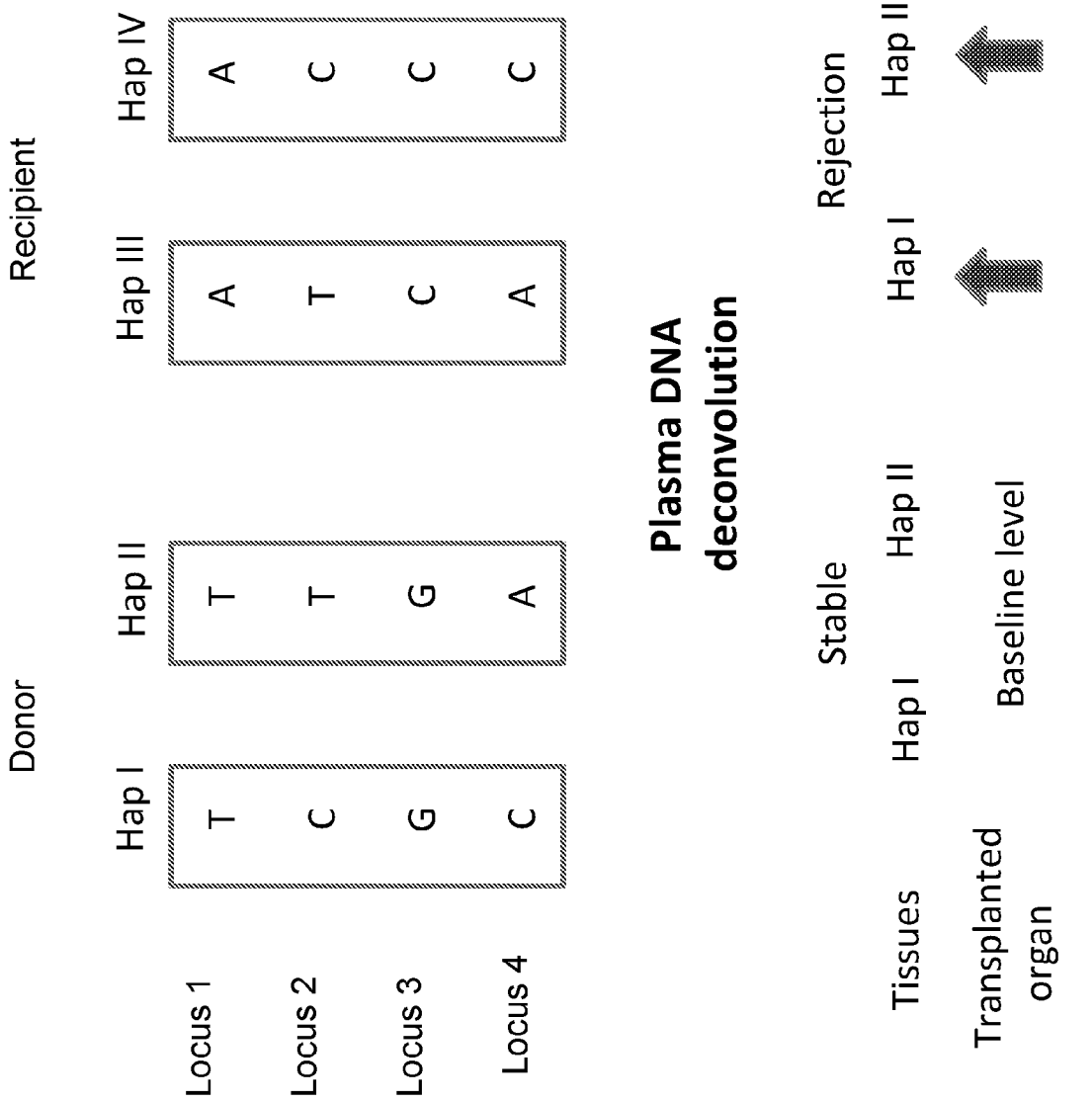

Donor
Recipient
Hap I
Hap II
Hap III
Hap IV
Locus 1
Locus 2
Locus 3
Locus 4
T
C
G
C
T
T
G
A
A
T
C
A
A
C
C
C
Plasma DNA deconvolution
Stable
Rejection
Tissues
Hap I
Hap II
Hap I
Hap II
Transplanted organ
Baseline level

FIG. 18

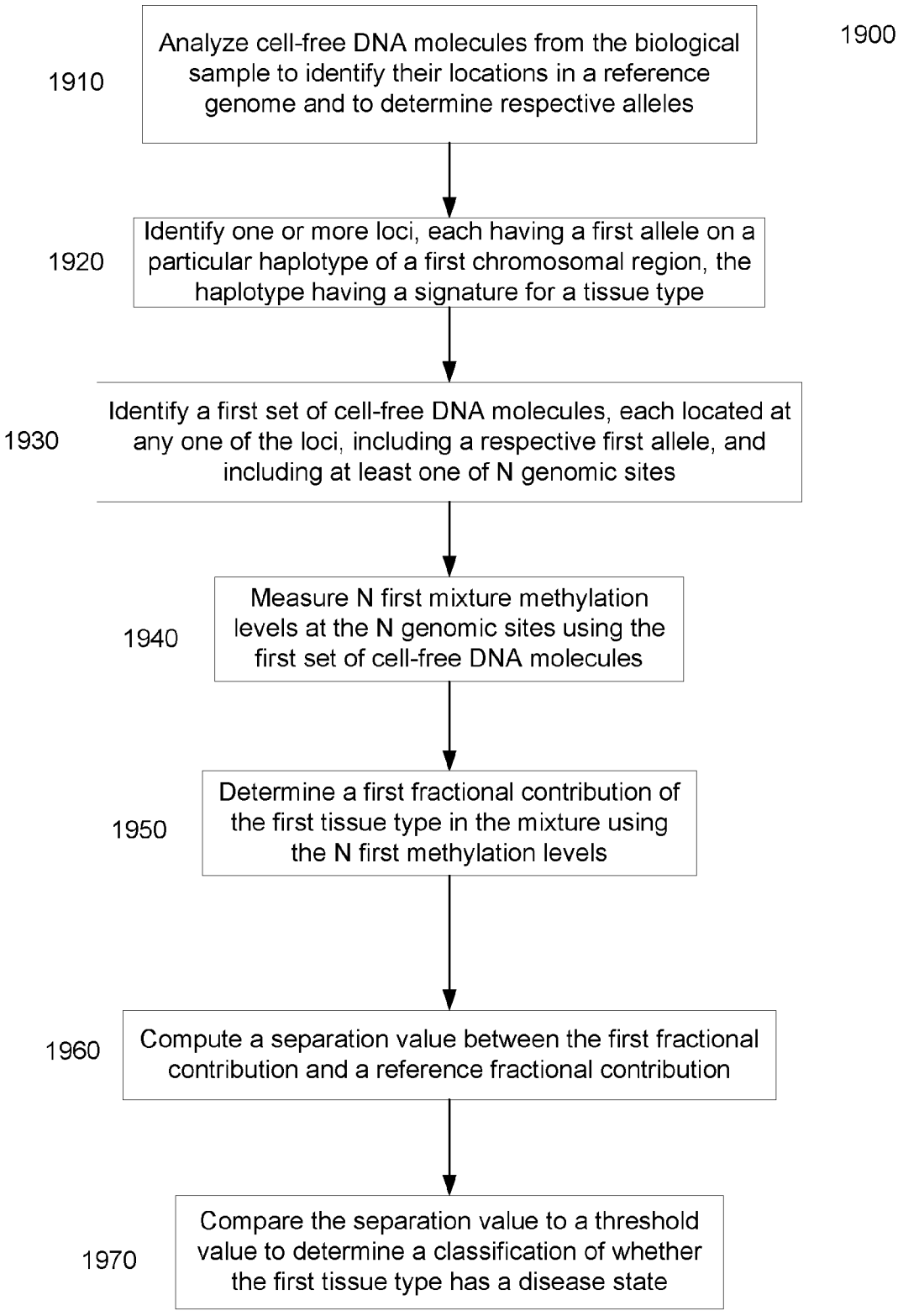

1900
1910
Analyze cell-free DNA molecules from the biological sample to identify their locations in a reference genome and to determine respective alleles
1920
Identify one or more loci, each having a first allele on a particular haplotype of a first chromosomal region, the haplotype having a signature for a tissue type
1930
Identify a first set of cell-free DNA molecules, each located at any one of the loci, including a respective first allele, and including at least one of N genomic sites
1940
Measure N first mixture methylation levels at the N genomic sites using the first set of cell-free DNA molecules
1950
Determine a first fractional contribution of the first tissue type in the mixture using the N first methylation levels
1960
Compute a separation value between the first fractional contribution and a reference fractional contribution
1970
Compare the separation value to a threshold value to determine a classification of whether the first tissue type has a disease state

FIG. 19

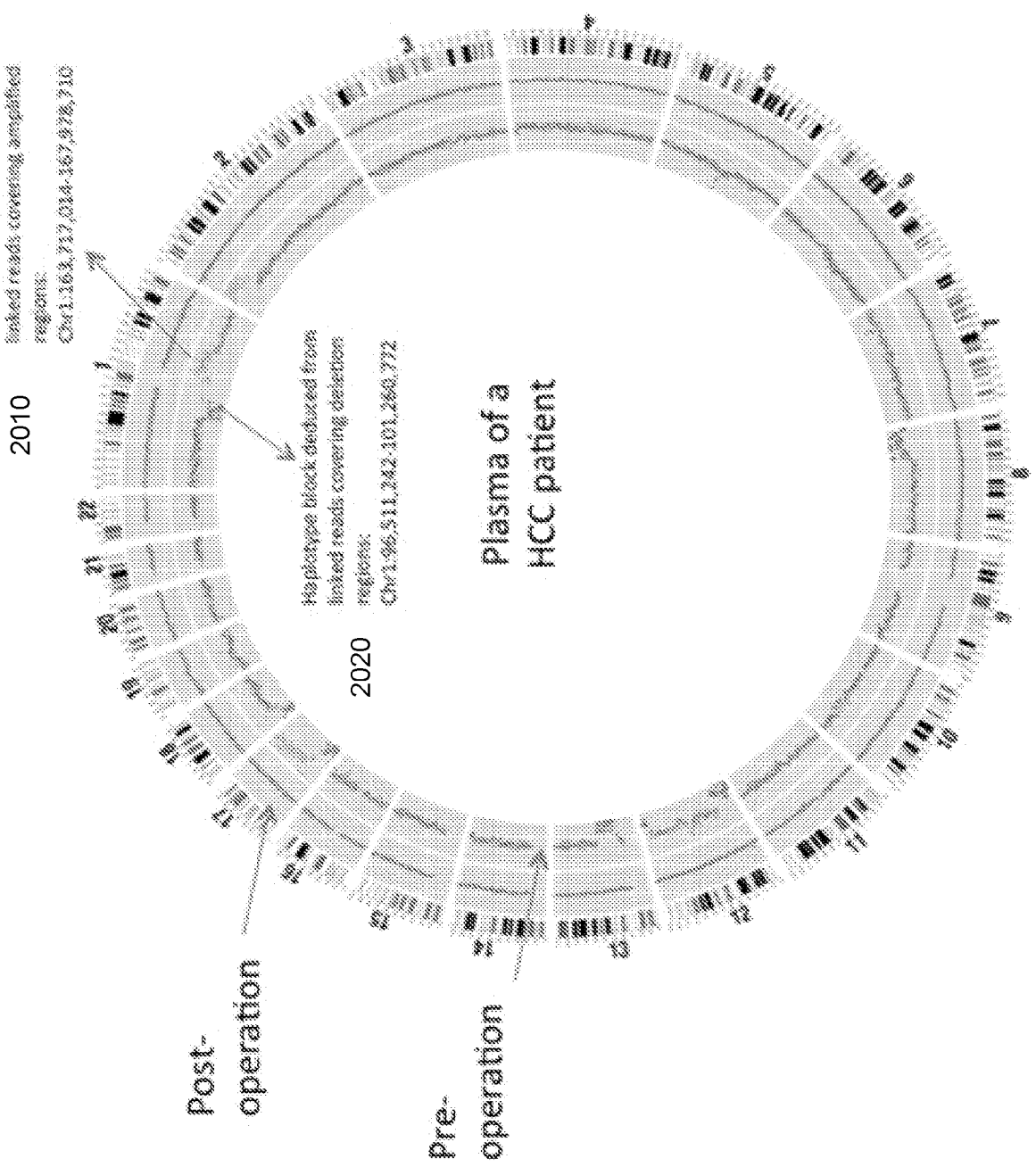

Plasma of a HCC patient
2010
2020
Haplotype block deduced from linked reads covering amplified region: Chr1:163,717,034-167,978,710
Haplotype block deduced from linked reads covering deletion region: Chr1:96,511,242-101,264,772
Post-operation
Pre-operation

FIG. 20

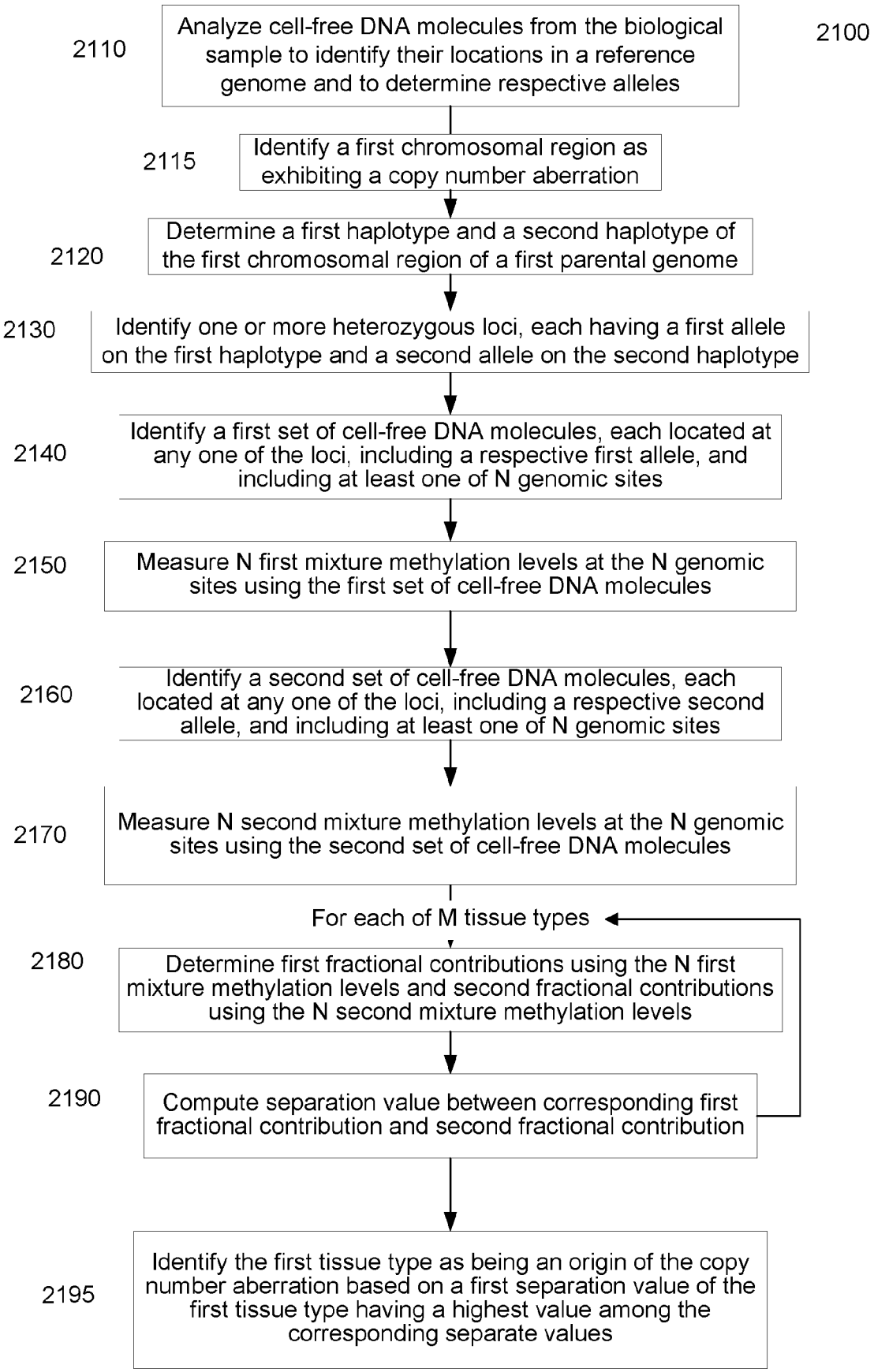

2100
2110
Analyze cell-free DNA molecules from the biological sample to identify their locations in a reference genome and to determine respective alleles
2115
Identify a first chromosomal region as exhibiting a copy number aberration
2120
Determine a first haplotype and a second haplotype of the first chromosomal region of a first parental genome
2130
Identify one or more heterozygous loci, each having a first allele on the first haplotype and a second allele on the second haplotype
2140
Identify a first set of cell-free DNA molecules, each located at any one of the loci, including a respective first allele, and including at least one of N genomic sites
2150
Measure N first mixture methylation levels at the N genomic sites using the first set of cell-free DNA molecules
2160
Identify a second set of cell-free DNA molecules, each located at any one of the loci, including a respective second allele, and including at least one of N genomic sites
2170
Measure N second mixture methylation levels at the N genomic sites using the second set of cell-free DNA molecules
For each of M tissue types
2180
Determine first fractional contributions using the N first mixture methylation levels and second fractional contributions using the N second mixture methylation levels
2190
Compute separation value between corresponding first fractional contribution and second fractional contribution
2195
Identify the first tissue type as being an origin of the copy number aberration based on a first separation value of the first tissue type having a highest value among the corresponding separate values

FIG. 21

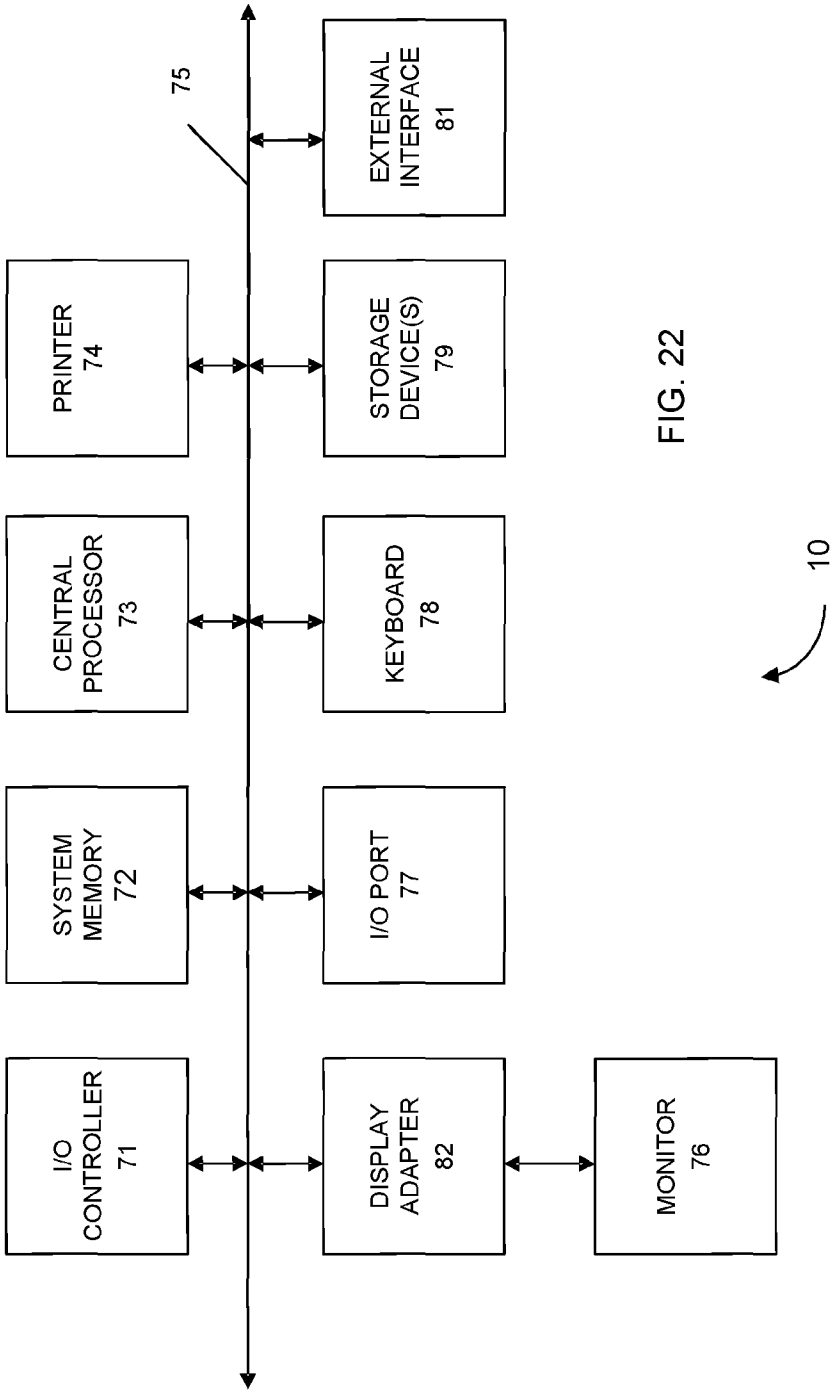

75
EXTERNAL INTERFACE 81
PRINTER 74
STORAGE DEVICE(S) 79
CENTRAL PROCESSOR 73
KEYBOARD 78
SYSTEM MEMORY 72
I/O PORT 77
I/O CONTROLLER 71
DISPLAY ADAPTER 82
MONITOR 76
10

FIG. 22

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2016/090642

A. CLASSIFICATION OF SUBJECT MATTER

C12Q 1/68(2006.01)i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

C12Q

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

DWPI,SIPOABS,CNABS,HKABS,CNKI,Pubmed,Google scholar:fetal, fetus, allele, haplotype, chromosome, parental, methylation, cell-free, non-invasive, THE CHINESE UNIVERSITY OF HONG KONG

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	WO 2014043763 A1 (UNIV HONG KONG CHINESEET AL.) 27 March 2014 (2014-03-27) see the whole document	1-65
A	WO 2011057094 A1 (UNIV HONG KONG CHINESEET AL.) 12 May 2011 (2011-05-12) see the whole document	1-65

☐ Further documents are listed in the continuation of Box C.

☑ See patent family annex.

* Special categories of cited documents:

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier application or patent but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
08 October 2016	**19 October 2016**

Name and mailing address of the ISA/CN	Authorized officer
STATE INTELLECTUAL PROPERTY OFFICE OF THE P.R.CHINA **6, Xitucheng Rd., Jimen Bridge, Haidian District, Beijing 100088** **China**	**LI,Kangqi**
Facsimile No. **(86-10)62019451**	Telephone No. **(86-10)62411034**

Form PCT/ISA/210 (second sheet) (July 2009)

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2016/090642

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)			Publication date (day/month/year)
WO	2014043763	A1	27 March 2014	JP	2015536639	A	24 December 2015
				AU	2013317708	A1	09 April 2015
				HK	1207124	A1	22 January 2016
				KR	20150082228	A	15 July 2015
				MX	2015003571	A	22 June 2015
				CA	2884066	A1	27 March 2014
				EP	2898100	A1	29 July 2015
				TW	201418474	A	16 May 2014
				EA	201500327	A1	30 July 2015
				SG	11201501927V	A	29 April 2015
				HK	1206797	A1	15 January 2016
				EP	2898100	A4	11 May 2016
				IL	237495	A	30 April 2015
				CN	104781422	A	15 July 2015
				PH	12015500547	A1	04 May 2015
				IL	237495	D0	30 April 2015
WO	2011057094	A1	12 May 2011	IL	219521	A	26 February 2015
				CN	105779280	A	20 July 2016
				IL	219521	D0	28 June 2012
				EP	2496717	A1	12 September 2012
				AU	2010315037	B2	18 September 2014
				US	2013253844	A1	26 September 2013
				US	8467976	B2	18 June 2013
				AU	2010315037	A8	12 July 2012
				JP	5540105	B2	02 July 2014
				CA	2779695	A1	12 May 2011
				CA	2779695	C	24 May 2016
				TW	I458976	B	01 November 2014
				MX	2012005214	A	21 September 2012
				AU	2010315037	B9	23 April 2015
				TW	201122470	A	01 July 2011
				US	2013323731	A1	05 December 2013
				JP	2014193165	A	09 October 2014
				EA	201200690	A1	30 May 2013
				CN	102770558	B	06 April 2016
				AU	2010315037	A1	07 June 2012
				CN	102770558	A	07 November 2012
				US	2011105353	A1	05 May 2011
				JP	2013509884	A	21 March 2013

Electronic Patent Application Fee Transmittal

Application Number:	16995607
Filing Date:	17-Aug-2020
Title of Invention:	DETERMINATION OF BASE MODIFICATIONS OF NUCLEIC ACIDS
First Named Inventor/Applicant Name:	Yuk-Ming Dennis Lo
Filer:	David Dah-Wei Hsu/Heather Orosco
Attorney Docket Number:	080015-028410US-1199353

Filed as Small Entity

Filing Fees for Utility under 35 USC 111(a)

Description	Fee Code	Quantity	Amount	Sub-Total in USD($)
Basic Filing:				
Pages:				
Claims:				
Miscellaneous-Filing:				
Petition:				
Patent-Appeals-and-Interference:				
Post-Allowance-and-Post-Issuance:				
Extension-of-Time:				

Description	Fee Code	Quantity	Amount	Sub-Total in USD($)
Miscellaneous:				
SUBMISSION- INFORMATION DISCLOSURE STMT	2806	1	130	130
		Total in USD ($)		**130**

Electronic Acknowledgement Receipt

EFS ID:	42463165
Application Number:	16995607
International Application Number:	
Confirmation Number:	1256
Title of Invention:	DETERMINATION OF BASE MODIFICATIONS OF NUCLEIC ACIDS
First Named Inventor/Applicant Name:	Yuk-Ming Dennis Lo
Customer Number:	20350
Filer:	David Dah-Wei Hsu/Heather Orosco
Filer Authorized By:	David Dah-Wei Hsu
Attorney Docket Number:	080015-028410US-1199353
Receipt Date:	15-APR-2021
Filing Date:	17-AUG-2020
Time Stamp:	13:50:53
Application Type:	Utility under 35 USC 111(a)

Payment information:

Submitted with Payment	yes
Payment Type	CARD
Payment was successfully received in RAM	$130
RAM confirmation Number	E20214ED52123946
Deposit Account	201430
Authorized User	Heather Orosco

The Director of the USPTO is hereby authorized to charge indicated fees and credit any overpayment as follows:

37 CFR 1.16 (National application filing, search, and examination fees)

37 CFR 1.17 (Patent application and reexamination processing fees)

37 CFR 1.19 (Document supply fees)

37 CFR 1.21 (Miscellaneous fees and charges)

File Listing:

Document Number	Document Description	File Name	File Size(Bytes)/ Message Digest	Multi Part /.zip	Pages (if appl.)
1		1199353_Supplemental_IDS.pdf	165745 57d7ee4a1dc02f8a99a1bd03ae2a9716944 3229a	yes	4
	Multipart Description/PDF files in .zip description				
	Document Description		**Start**	**End**	
	Transmittal Letter		1	2	
	Information Disclosure Statement (IDS) Form (SB08)		3	4	
Warnings:					
Information:					
2	Foreign Reference	REF_WO2017012544A1.pdf	6088361 ae26984903a769b593a2993cdc71947424c d420e	no	127
Warnings:					
Information:					
3	Non Patent Literature	REF_Combined_Search_and_Exam_Report_for_GB2103003_6_dated_2021-03-29.pdf	191749 d4ef194a5ec95b817ca53abe9c165b5103d 11870	no	5
Warnings:					
Information:					
4	Fee Worksheet (SB06)	fee-info.pdf	30416 2536dc44eac7d7e617a13b669a8156b629d abead	no	2
Warnings:					
Information:					
		Total Files Size (in bytes):	6476271		

This Acknowledgement Receipt evidences receipt on the noted date by the USPTO of the indicated documents, characterized by the applicant, and including page counts, where applicable. It serves as evidence of receipt similar to a Post Card, as described in MPEP 503.

New Applications Under 35 U.S.C. 111

If a new application is being filed and the application includes the necessary components for a filing date (see 37 CFR 1.53(b)-(d) and MPEP 506), a Filing Receipt (37 CFR 1.54) will be issued in due course and the date shown on this Acknowledgement Receipt will establish the filing date of the application.

National Stage of an International Application under 35 U.S.C. 371

If a timely submission to enter the national stage of an international application is compliant with the conditions of 35 U.S.C. 371 and other applicable requirements a Form PCT/DO/EO/903 indicating acceptance of the application as a national stage submission under 35 U.S.C. 371 will be issued in addition to the Filing Receipt, in due course.

New International Application Filed with the USPTO as a Receiving Office

If a new international application is being filed and the international application includes the necessary components for an international filing date (see PCT Article 11 and MPEP 1810), a Notification of the International Application Number and of the International Filing Date (Form PCT/RO/105) will be issued in due course, subject to prescriptions concerning national security, and the date shown on this Acknowledgement Receipt will establish the international filing date of the application.

I hereby certify that this correspondence is being filed via EFS-Web with the United States Patent and Trademark Office on April 15, 2021.

KILPATRICK TOWNSEND & STOCKTON LLP

By: /Heather A. Orosco/
Heather A. Orosco

PATENT
Attorney Docket No.: 080015-028410US-1199353
Client Ref. No.: 19/MED/901 US Reg

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

In re Application of:	Confirmation No.: 1256
Yuk-Ming Dennis Lo	Examiner: Laura B. Goddard
Application No.: 16/995,607	Technology Center/Art Unit: 1642
Filed: August 17, 2020	
For: DETERMINATION OF BASE MODIFICATIONS OF NUCLEIC ACIDS	SUPPLEMENTAL INFORMATION DISCLOSURE STATEMENT
Customer No.: 20350	

Commissioner for Patents
P.O. Box 1450
Alexandria, VA 22313-1450

Commissioner:

The references cited on attached form PTO/SB/08A are being called to the attention of the Examiner. In accordance with the provisions of 37 CFR §1.98(a)(2), copies of any cited U.S. Patent and U.S. Patent Application Publications are not provided. Copies of the remaining cited references are provided.

It is respectfully requested that the cited references be expressly considered during the prosecution of this application, and the references be made of record therein and appear among the "references cited" on any patent to issue therefrom.

As provided for by 37 CFR §1.97(g) and (h), no inference should be made that the information and references cited are prior art merely because they are in this statement and no

representation is being made that a search has been conducted or that this statement encompasses all the possible relevant information.

This IDS is being filed after an action closing prosecution, but before the payment of the Issue Fee.

I hereby certify under 37 CFR §1.97(e)(1) that each item of information contained in this Information Disclosure Statement was first cited in any communication from a foreign patent office in a counterpart foreign application not more than three months prior to the filing of this statement.

The Commissioner is authorized to charge the fee set forth in §1.17(p) to the firm's credit card for consideration of this paper. The Commissioner is authorized to charge any additional fee due to Deposit Account No. 20-1430.

Respectfully submitted,

/David D. Hsu/

David D. Hsu
Registration No. 71659

KILPATRICK TOWNSEND & STOCKTON LLP

Attachment

74687977V.1

UNITED STATES PATENT AND TRADEMARK OFFICE

UNITED STATES DEPARTMENT OF COMMERCE
United States Patent and Trademark Office
Address: COMMISSIONER FOR PATENTS
P.O. Box 1450
Alexandria, Virginia 22313-1450
www.uspto.gov

NOTICE OF ALLOWANCE AND FEE(S) DUE

20350 7590 04/07/2021

Kilpatrick Townsend & Stockton LLP - West Coast
Mailstop: IP Docketing - 22
1100 Peachtree Street
Suite 2800
Atlanta, GA 30309

EXAMINER
GODDARD, LAURA B

ART UNIT	PAPER NUMBER
1642	

DATE MAILED: 04/07/2021

APPLICATION NO.	FILING DATE	FIRST NAMED INVENTOR	ATTORNEY DOCKET NO.	CONFIRMATION NO.
16/995,607	08/17/2020	Yuk-Ming Dennis Lo	080015-028410US-1199353	1256

TITLE OF INVENTION: DETERMINATION OF BASE MODIFICATIONS OF NUCLEIC ACIDS

APPLN. TYPE	ENTITY STATUS	ISSUE FEE DUE	PUBLICATION FEE DUE	PREV. PAID ISSUE FEE	TOTAL FEE(S) DUE	DATE DUE
nonprovisional	SMALL	$600	$0.00	$0.00	$600	07/07/2021

THE APPLICATION IDENTIFIED ABOVE HAS BEEN EXAMINED AND IS ALLOWED FOR ISSUANCE AS A PATENT. PROSECUTION ON THE MERITS IS CLOSED. THIS NOTICE OF ALLOWANCE IS NOT A GRANT OF PATENT RIGHTS. THIS APPLICATION IS SUBJECT TO WITHDRAWAL FROM ISSUE AT THE INITIATIVE OF THE OFFICE OR UPON PETITION BY THE APPLICANT. SEE 37 CFR 1.313 AND MPEP 1308.

THE ISSUE FEE AND PUBLICATION FEE (IF REQUIRED) MUST BE PAID WITHIN THREE MONTHS FROM THE MAILING DATE OF THIS NOTICE OR THIS APPLICATION SHALL BE REGARDED AS ABANDONED. THIS STATUTORY PERIOD CANNOT BE EXTENDED. SEE 35 U.S.C. 151. THE ISSUE FEE DUE INDICATED ABOVE DOES NOT REFLECT A CREDIT FOR ANY PREVIOUSLY PAID ISSUE FEE IN THIS APPLICATION. IF AN ISSUE FEE HAS PREVIOUSLY BEEN PAID IN THIS APPLICATION (AS SHOWN ABOVE), THE RETURN OF PART B OF THIS FORM WILL BE CONSIDERED A REQUEST TO REAPPLY THE PREVIOUSLY PAID ISSUE FEE TOWARD THE ISSUE FEE NOW DUE.

HOW TO REPLY TO THIS NOTICE:

I. Review the ENTITY STATUS shown above. If the ENTITY STATUS is shown as SMALL or MICRO, verify whether entitlement to that entity status still applies.

If the ENTITY STATUS is the same as shown above, pay the TOTAL FEE(S) DUE shown above.

If the ENTITY STATUS is changed from that shown above, on PART B - FEE(S) TRANSMITTAL, complete section number 5 titled "Change in Entity Status (from status indicated above)".

For purposes of this notice, small entity fees are 1/2 the amount of undiscounted fees, and micro entity fees are 1/2 the amount of small entity fees.

II. PART B - FEE(S) TRANSMITTAL, or its equivalent, must be completed and returned to the United States Patent and Trademark Office (USPTO) with your ISSUE FEE and PUBLICATION FEE (if required). If you are charging the fee(s) to your deposit account, section "4b" of Part B - Fee(s) Transmittal should be completed and an extra copy of the form should be submitted. If an equivalent of Part B is filed, a request to reapply a previously paid issue fee must be clearly made, and delays in processing may occur due to the difficulty in recognizing the paper as an equivalent of Part B.

III. All communications regarding this application must give the application number. Please direct all communications prior to issuance to Mail Stop ISSUE FEE unless advised to the contrary.

IMPORTANT REMINDER: Maintenance fees are due in utility patents issuing on applications filed on or after Dec. 12, 1980. It is patentee's responsibility to ensure timely payment of maintenance fees when due. More information is available at www.uspto.gov/PatentMaintenanceFees.

PART B - FEE(S) TRANSMITTAL

Complete and send this form, together with applicable fee(s), by mail or fax, or via EFS-Web.

By mail, send to: Mail Stop ISSUE FEE
Commissioner for Patents
P.O. Box 1450
Alexandria, Virginia 22313-1450

By fax, send to: (571)-273-2885

INSTRUCTIONS: This form should be used for transmitting the ISSUE FEE and PUBLICATION FEE (if required). Blocks 1 through 5 should be completed where appropriate. All further correspondence including the Patent, advance orders and notification of maintenance fees will be mailed to the current correspondence address as indicated unless corrected below or directed otherwise in Block 1, by (a) specifying a new correspondence address; and/or (b) indicating a separate "FEE ADDRESS" for maintenance fee notifications.

CURRENT CORRESPONDENCE ADDRESS (Note: Use Block 1 for any change of address)

20350 7590 04/07/2021

Kilpatrick Townsend & Stockton LLP - West Coast
Mailstop: IP Docketing - 22
1100 Peachtree Street
Suite 2800
Atlanta, GA 30309

Note: A certificate of mailing can only be used for domestic mailings of the Fee(s) Transmittal. This certificate cannot be used for any other accompanying papers. Each additional paper, such as an assignment or formal drawing, must have its own certificate of mailing or transmission.

Certificate of Mailing or Transmission

I hereby certify that this Fee(s) Transmittal is being deposited with the United States Postal Service with sufficient postage for first class mail in an envelope addressed to the Mail Stop ISSUE FEE address above, or being transmitted to the USPTO via EFS-Web or by facsimile to (571) 273-2885, on the date below.

(Typed or printed name)
(Signature)
(Date)

APPLICATION NO.	FILING DATE	FIRST NAMED INVENTOR	ATTORNEY DOCKET NO.	CONFIRMATION NO.
16/995,607	08/17/2020	Yuk-Ming Dennis Lo	080015-028410US-1199353	1256

TITLE OF INVENTION: DETERMINATION OF BASE MODIFICATIONS OF NUCLEIC ACIDS

APPLN. TYPE	ENTITY STATUS	ISSUE FEE DUE	PUBLICATION FEE DUE	PREV. PAID ISSUE FEE	TOTAL FEE(S) DUE	DATE DUE
nonprovisional	SMALL	$600	$0.00	$0.00	$600	07/07/2021

EXAMINER	ART UNIT	CLASS-SUBCLASS
GODDARD, LAURA B	1642	435-006110

1. Change of correspondence address or indication of "Fee Address" (37 CFR 1.363).

❑ Change of correspondence address (or Change of Correspondence Address form PTO/SB/122) attached.

❑ "Fee Address" indication (or "Fee Address" Indication form PTO/SB/47; Rev 03-09 or more recent) attached. **Use of a Customer Number is required.**

2. For printing on the patent front page, list
(1) The names of up to 3 registered patent attorneys or agents OR, alternatively,
(2) The name of a single firm (having as a member a registered attorney or agent) and the names of up to 2 registered patent attorneys or agents. If no name is listed, no name will be printed.

1 _____
2 _____
3 _____

3. ASSIGNEE NAME AND RESIDENCE DATA TO BE PRINTED ON THE PATENT (print or type)

PLEASE NOTE: Unless an assignee is identified below, no assignee data will appear on the patent. If an assignee is identified below, the document must have been previously recorded, or filed for recordation, as set forth in 37 CFR 3.11 and 37 CFR 3.81(a). Completion of this form is NOT a substitute for filing an assignment.

(A) NAME OF ASSIGNEE (B) RESIDENCE: (CITY and STATE OR COUNTRY)

Please check the appropriate assignee category or categories (will not be printed on the patent) : ❑ Individual ❑ Corporation or other private group entity ❑ Government

4a. Fees submitted: ❑ Issue Fee ❑ Publication Fee (if required) ❑ Advance Order - # of Copies _____

4b. Method of Payment: *(Please first reapply any previously paid fee shown above)*

❑ Electronic Payment via EFS-Web ❑ Enclosed check ❑ Non-electronic payment by credit card (Attach form PTO-2038)

❑ The Director is hereby authorized to charge the required fee(s), any deficiency, or credit any overpayment to Deposit Account No. _____

5. **Change in Entity Status** (from status indicated above)

❑ Applicant certifying micro entity status. See 37 CFR 1.29

❑ Applicant asserting small entity status. See 37 CFR 1.27

❑ Applicant changing to regular undiscounted fee status.

NOTE: Absent a valid certification of Micro Entity Status (see forms PTO/SB/15A and 15B), issue fee payment in the micro entity amount will not be accepted at the risk of application abandonment.
NOTE: If the application was previously under micro entity status, checking this box will be taken to be a notification of loss of entitlement to micro entity status.
NOTE: Checking this box will be taken to be a notification of loss of entitlement to small or micro entity status, as applicable.

NOTE: This form must be signed in accordance with 37 CFR 1.31 and 1.33. See 37 CFR 1.4 for signature requirements and certifications.

Authorized Signature _____ Date _____

Typed or printed name _____ Registration No. _____

UNITED STATES PATENT AND TRADEMARK OFFICE

UNITED STATES DEPARTMENT OF COMMERCE
United States Patent and Trademark Office
Address: COMMISSIONER FOR PATENTS
P.O. Box 1450
Alexandria, Virginia 22313-1450
www.uspto.gov

APPLICATION NO.	FILING DATE	FIRST NAMED INVENTOR	ATTORNEY DOCKET NO.	CONFIRMATION NO.
16/995,607	08/17/2020	Yuk-Ming Dennis Lo	080015-028410US-1199353	1256

20350 7590 04/07/2021

Kilpatrick Townsend & Stockton LLP - West Coast
Mailstop: IP Docketing - 22
1100 Peachtree Street
Suite 2800
Atlanta, GA 30309

EXAMINER	
GODDARD, LAURA B	

ART UNIT	PAPER NUMBER
1642	

DATE MAILED: 04/07/2021

Determination of Patent Term Adjustment under 35 U.S.C. 154 (b)
(Applications filed on or after May 29, 2000)

The Office has discontinued providing a Patent Term Adjustment (PTA) calculation with the Notice of Allowance.

Section 1(h)(2) of the AIA Technical Corrections Act amended 35 U.S.C. 154(b)(3)(B)(i) to eliminate the requirement that the Office provide a patent term adjustment determination with the notice of allowance. See Revisions to Patent Term Adjustment, 78 Fed. Reg. 19416, 19417 (Apr. 1, 2013). Therefore, the Office is no longer providing an initial patent term adjustment determination with the notice of allowance. The Office will continue to provide a patent term adjustment determination with the Issue Notification Letter that is mailed to applicant approximately three weeks prior to the issue date of the patent, and will include the patent term adjustment on the patent. Any request for reconsideration of the patent term adjustment determination (or reinstatement of patent term adjustment) should follow the process outlined in 37 CFR 1.705.

Any questions regarding the Patent Term Extension or Adjustment determination should be directed to the Office of Patent Legal Administration at (571)-272-7702. Questions relating to issue and publication fee payments should be directed to the Customer Service Center of the Office of Patent Publication at 1-(888)-786-0101 or (571)-272-4200.

OMB Clearance and PRA Burden Statement for PTOL-85 Part B

The Paperwork Reduction Act (PRA) of 1995 requires Federal agencies to obtain Office of Management and Budget approval before requesting most types of information from the public. When OMB approves an agency request to collect information from the public, OMB (i) provides a valid OMB Control Number and expiration date for the agency to display on the instrument that will be used to collect the information and (ii) requires the agency to inform the public about the OMB Control Number's legal significance in accordance with 5 CFR 1320.5(b).

The information collected by PTOL-85 Part B is required by 37 CFR 1.311. The information is required to obtain or retain a benefit by the public which is to file (and by the USPTO to process) an application. Confidentiality is governed by 35 U.S.C. 122 and 37 CFR 1.14. This collection is estimated to take 30 minutes to complete, including gathering, preparing, and submitting the completed application form to the USPTO. Time will vary depending upon the individual case. Any comments on the amount of time you require to complete this form and/or suggestions for reducing this burden, should be sent to the Chief Information Officer, U.S. Patent and Trademark Office, U.S. Department of Commerce, P.O. Box 1450, Alexandria, Virginia 22313-1450. DO NOT SEND FEES OR COMPLETED FORMS TO THIS ADDRESS. SEND TO: Commissioner for Patents, P.O. Box 1450, Alexandria, Virginia 22313-1450. Under the Paperwork Reduction Act of 1995, no persons are required to respond to a collection of information unless it displays a valid OMB control number.

Privacy Act Statement

The Privacy Act of 1974 (P.L. 93-579) requires that you be given certain information in connection with your submission of the attached form related to a patent application or patent. Accordingly, pursuant to the requirements of the Act, please be advised that: (1) the general authority for the collection of this information is 35 U.S.C. 2(b)(2); (2) furnishing of the information solicited is voluntary; and (3) the principal purpose for which the information is used by the U.S. Patent and Trademark Office is to process and/or examine your submission related to a patent application or patent. If you do not furnish the requested information, the U.S. Patent and Trademark Office may not be able to process and/or examine your submission, which may result in termination of proceedings or abandonment of the application or expiration of the patent.

The information provided by you in this form will be subject to the following routine uses:

1. The information on this form will be treated confidentially to the extent allowed under the Freedom of Information Act (5 U.S.C. 552) and the Privacy Act (5 U.S.C 552a). Records from this system of records may be disclosed to the Department of Justice to determine whether disclosure of these records is required by the Freedom of Information Act.
2. A record from this system of records may be disclosed, as a routine use, in the course of presenting evidence to a court, magistrate, or administrative tribunal, including disclosures to opposing counsel in the course of settlement negotiations.
3. A record in this system of records may be disclosed, as a routine use, to a Member of Congress submitting a request involving an individual, to whom the record pertains, when the individual has requested assistance from the Member with respect to the subject matter of the record.
4. A record in this system of records may be disclosed, as a routine use, to a contractor of the Agency having need for the information in order to perform a contract. Recipients of information shall be required to comply with the requirements of the Privacy Act of 1974, as amended, pursuant to 5 U.S.C. 552a(m).
5. A record related to an International Application filed under the Patent Cooperation Treaty in this system of records may be disclosed, as a routine use, to the International Bureau of the World Intellectual Property Organization, pursuant to the Patent Cooperation Treaty.
6. A record in this system of records may be disclosed, as a routine use, to another federal agency for purposes of National Security review (35 U.S.C. 181) and for review pursuant to the Atomic Energy Act (42 U.S.C. 218(c)).
7. A record from this system of records may be disclosed, as a routine use, to the Administrator, General Services, or his/her designee, during an inspection of records conducted by GSA as part of that agency's responsibility to recommend improvements in records management practices and programs, under authority of 44 U.S.C. 2904 and 2906. Such disclosure shall be made in accordance with the GSA regulations governing inspection of records for this purpose, and any other relevant (i.e., GSA or Commerce) directive. Such disclosure shall not be used to make determinations about individuals.
8. A record from this system of records may be disclosed, as a routine use, to the public after either publication of the application pursuant to 35 U.S.C. 122(b) or issuance of a patent pursuant to 35 U.S.C. 151. Further, a record may be disclosed, subject to the limitations of 37 CFR 1.14, as a routine use, to the public if the record was filed in an application which became abandoned or in which the proceedings were terminated and which application is referenced by either a published application, an application open to public inspection or an issued patent.
9. A record from this system of records may be disclosed, as a routine use, to a Federal, State, or local law enforcement agency, if the USPTO becomes aware of a violation or potential violation of law or regulation.

Notice of Allowability	Application No. 16/995,607	Applicant(s) Lo et al.	
	Examiner LAURA B GODDARD	Art Unit 1642	AIA (FITF) Status Yes

-- The MAILING DATE of this communication appears on the cover sheet with the correspondence address--

All claims being allowable, PROSECUTION ON THE MERITS IS (OR REMAINS) CLOSED in this application. If not included herewith (or previously mailed), a Notice of Allowance (PTOL-85) or other appropriate communication will be mailed in due course. **THIS NOTICE OF ALLOWABILITY IS NOT A GRANT OF PATENT RIGHTS.** This application is subject to withdrawal from issue at the initiative of the Office or upon petition by the applicant. See 37 CFR 1.313 and MPEP 1308.

1. ☑ This communication is responsive to 3/12/2021.
   ☐ A declaration(s)/affidavit(s) under **37 CFR 1.130(b)** was/were filed on _____.

2. ☐ An election was made by the applicant in response to a restriction requirement set forth during the interview on _____; the restriction requirement and election have been incorporated into this action.

3. ☑ The allowed claim(s) is/are 1-17 and 70-71 . As a result of the allowed claim(s), you may be eligible to benefit from the **Patent Prosecution Highway** program at a participating intellectual property office for the corresponding application. For more information , please see **http://www.uspto.gov/patents/init_events/pph/index.jsp** or send an inquiry to **PPHfeedback@uspto.gov.**

4. ☐ Acknowledgment is made of a claim for foreign priority under 35 U.S.C. § 119(a)-(d) or (f).

   Certified copies:

   a) ☐ All b) ☐ Some *c) ☐ None of the:

   1. ☐ Certified copies of the priority documents have been received.
   2. ☐ Certified copies of the priority documents have been received in Application No. _____ .
   3. ☐ Copies of the certified copies of the priority documents have been received in this national stage application from the International Bureau (PCT Rule 17.2(a)).

   * Certified copies not received: _____ .

Applicant has THREE MONTHS FROM THE "MAILING DATE" of this communication to file a reply complying with the requirements noted below. Failure to timely comply will result in ABANDONMENT of this application.
THIS THREE-MONTH PERIOD IS NOT EXTENDABLE.

5. ☐ CORRECTED DRAWINGS (as "replacement sheets") must be submitted.
   ☐ including changes required by the attached Examiner's Amendment / Comment or in the Office action of Paper No./Mail Date _____ .

   Identifying indicia such as the application number (see 37 CFR 1.84(c)) should be written on the drawings in the front (not the back) of each sheet. Replacement sheet(s) should be labeled as such in the header according to 37 CFR 1.121(d).

6. ☐ DEPOSIT OF and/or INFORMATION about the deposit of BIOLOGICAL MATERIAL must be submitted. Note the attached Examiner's comment regarding REQUIREMENT FOR THE DEPOSIT OF BIOLOGICAL MATERIAL.

Attachment(s)

1. ☐ Notice of References Cited (PTO-892)
2. ☑ Information Disclosure Statements (PTO/SB/08), Paper No./Mail Date _____.
3. ☐ Examiner's Comment Regarding Requirement for Deposit of Biological Material _____.
4. ☐ Interview Summary (PTO-413), Paper No./Mail Date. _____.
5. ☐ Examiner's Amendment/Comment
6. ☐ Examiner's Statement of Reasons for Allowance
7. ☐ Other _____.

/Laura B Goddard/
Primary Examiner, Art Unit 1642

Search Notes	Application/Control No.	Applicant(s)/Patent Under Reexamination
	16/995,607	Lo et al.
	Examiner	Art Unit
	LAURA B GODDARD	1642

CPC - Searched*		
Symbol	Date	Examiner

CPC Combination Sets - Searched*		
Symbol	Date	Examiner
C12Q1/6816 or C12N9/22 or G16B20/00 or G16B30/00 or C12Q1/6851	03/29/2021	LBG

US Classification - Searched*			
Class	Subclass	Date	Examiner

* See search history printout included with this form or the SEARCH NOTES box below to determine the scope of the search.

Search Notes		
Search Notes	Date	Examiner
WEST	12/23/2020	LBG
Google	12/23/2020	LBG
STN	12/23/2020	LBG
WEST	03/23/2021	LBG
101 expert consult with SPE Daniel Kolker	03/29/2021	LBG
WEST	03/29/2021	LBG
Google and Google Scholar	03/29/2021	LBG

/Laura B Goddard/ Primary Examiner, Art Unit 1642	

Search Notes	Application/Control No.	Applicant(s)/Patent Under Reexamination
	16/995,607	Lo et al.
	Examiner	Art Unit
	LAURA B GODDARD	1642

Interference Search			
US Class/CPC Symbol	US Subclass/CPC Group	Date	Examiner
C12Q	1/6816	03/29/2021	LBG
G16B	30/00	03/29/2021	LBG
G16B	20/00	03/29/2021	LBG

/Laura B Goddard/ Primary Examiner, Art Unit 1642	

Issue Classification	Application/Control No.	Applicant(s)/Patent Under Reexamination
	16/995,607	Lo et al.
	Examiner	Art Unit
	LAURA B GODDARD	1642

CPC						
Symbol					Type	Version
C12Q	/	1	/	6816	F	2013-01-01
C12N	/	9	/	22	I	2013-01-01
G16B	/	20	/	00	I	2019-02-01
G16B	/	30	/	00	I	2019-02-01
C12Q	/	1	/	6851	I	2013-01-01

CPC Combination Sets								
Symbol					Type	Set	Ranking	Version
	/		/					

NONE		Total Claims Allowed:	
(Assistant Examiner)	(Date)	19	
/Laura B Goddard/ Primary Examiner, Art Unit 1642	29 March 2021	O.G. Print Claim(s)	O.G. Print Figure
(Primary Examiner)	(Date)	1	none

U.S. Patent and Trademark Office

Part of Paper No.: 20210329

Issue Classification	Application/Control No.	Applicant(s)/Patent Under Reexamination
	16/995,607	Lo et al.
	Examiner	Art Unit
	LAURA B GODDARD	1642

INTERNATIONAL CLASSIFICATION				
CLAIMED				
C12Q	/	1	/	6816
G16B	/	30	/	00
G16B	/	20	/	00
	/		/	

NON-CLAIMED				
	/		/	

US ORIGINAL CLASSIFICATION	
CLASS	SUBCLASS

CROSS REFERENCES(S)						
CLASS	SUBCLASS (ONE SUBCLASS PER BLOCK)					

NONE (Assistant Examiner)	(Date)	Total Claims Allowed: 19	
/Laura B Goddard/ Primary Examiner, Art Unit 1642 (Primary Examiner)	29 March 2021 (Date)	O.G. Print Claim(s) 1	O.G. Print Figure none

Issue Classification	Application/Control No.	Applicant(s)/Patent Under Reexamination
	16/995,607	Lo et al.
	Examiner	Art Unit
	LAURA B GODDARD	1642

☑ Claims renumbered in the same order as presented by applicant ☐ CPA ☐ T.D. ☐ R.1.47

CLAIMS

Final	Original	Final	Original	Final	Original	Final	Original	Final	Original	Final	Original	Final	Original	Final	Original

NONE		Total Claims Allowed:	
(Assistant Examiner)	(Date)	19	
/Laura B Goddard/ Primary Examiner, Art Unit 1642	29 March 2021	O.G. Print Claim(s)	O.G. Print Figure
(Primary Examiner)	(Date)	1	none

U.S. Patent and Trademark Office

Part of Paper No.: 20210329

INFORMATION DISCLOSURE STATEMENT BY APPLICANT (Not for submission under 37 CFR 1.99)		
	Application Number	16/995,607
	Filing Date	August 17, 2020
	First Named Inventor	Yuk-Ming Dennis Lo
	Art Unit	1642
	Examiner Name	Laura B. Goddard
	Attorney Docket Number	080015-028410US-1199353

U.S. PATENTS						
Examiner Initial*	Cite No	Patent Number	Kind Code [1]	Issue Date	Name of Patentee or Applicant of cited Document	Pages, Columns, Lines, Where Relevant Passages or Relevant Figures Appear

U.S. PATENT APPLICATION PUBLICATIONS						
Examiner Initial*	Cite No	Publication Number	Kind Code [1]	Publication Date	Name of Patentee or Applicant of cited Document	Pages, Columns, Lines, Where Relevant Passages or Relevant Figures Appear
	1.	2013/0230909	A1	09-05-2013	Pan et al.	

FOREIGN PATENT DOCUMENTS								
Examiner Initial*	Cite No	Foreign Document Number[3]	Country Code[2]	Kind Code[4]	Publication Date	Name of Patentee or Applicant of cited Document	Pages, Columns, Lines where Relevant Passages or Relevant Figures Appear	T[5]
	2.	2010068289	WO	A2	06-17-2010	Pacific Biosciences of California, Inc.		
	3.	2010027484	WO	A2	03-11-2010	Pacific Biosciences of California, Inc.		
	4.	2014153757	WO	A1	10-02-2014	BGI Shenzhen	English machine translation attached.	
	5.	2013188846	WO	A1	12-19-2013	Harry Stylli		
	6.	104053789	CN	A	09-17-2014	BGI Health Service Co., Ltd.	English abstract attached. Equivalent to AU 2012380221 B2 below.	
	7.	104053789	CN	B	02-10-2016	BGI Health Service Co., Ltd.	English abstract attached. Equivalent to AU 2012380221 B2 below.	
	8.	2013170429	WO	A1	11-21-2013	BGI Health Service Co., Ltd.	English abstract attached. Equivalent to AU 2012380221 B2 below.	

ALL REFERENCES CONSIDERED EXCEPT WHERE LINED THROUGH. /LBG/

INFORMATION DISCLOSURE STATEMENT BY APPLICANT (Not for submission under 37 CFR 1.99)		
	Application Number	16/995,607
	Filing Date	August 17, 2020
	First Named Inventor	Yuk-Ming Dennis Lo
	Art Unit	1642
	Examiner Name	Laura B. Goddard
	Attorney Docket Number	080015-028410US-1199353

	9.	2012380221	AU	B2	09-29-2016	BGI Diagnosis Co., Ltd.	Equivalent to CN 104053789 A & B and to WO 2013170429 A1 above.	

NON-PATENT LITERATURE DOCUMENTS			
Examiner Initials*	Cite No	Include name of the author (in CAPITAL LETTERS), title of the article (when appropriate), title of the item (book, magazine, journal, serial, symposium, catalog, etc.), date, page(s), volume-issue number(s), publisher, city and/or country where published.	T[5]
	10.	International Search Report and Written Opinion mailed November 19, 2020 in International Patent Application No. PCT/CN2020/109602. 11 pages.	

EXAMINER SIGNATURE			
Examiner Signature	/Laura B Goddard/	Date Considered	03/23/2021

*EXAMINER: Initial if reference considered, whether or not citation is in conformance with MPEP 609. Draw line through a citation if not in conformance and not considered. Include copy of this form with next communication to applicant.

[1] See Kind Codes of USPTO Patent Documents at www.USPTO.GOV or MPEP 901.04. [2] Enter office that issued the document, by the two-letter code (WIPO Standard ST.3). [3] For Japanese patent documents, the indication of the year of the reign of the Emperor must precede the serial number of the patent document.
[4] Kind of document by the appropriate symbols as indicated on the document under WIPO Standard ST.16 if possible. [5] Applicant is to place a check mark here if English language translation is attached.

ALL REFERENCES CONSIDERED EXCEPT WHERE LINED THROUGH. /LBG/

Web Search History

date, time	web site	search string
3/29/2021 12:42:01 PM	Google	sequencing methylation detection pulse width duration
3/29/2021 12:55:01 PM	Google	sequencing methylation detection pulse width duration software
3/29/2021 1:09:20 PM	Google Scholar	sequencing methylation detection pulse width duration software
3/29/2021 1:10:27 PM	Google Scholar	sequencing methylation detection pulse width duration software [before:2020]
3/29/2021 1:35:19 PM	Google Scholar	[before:2020] SMRT sequencing methylation detection pulse width duration software
3/29/2021 1:36:09 PM	Google Scholar	SMRT sequencing methylation detection pulse width duration software [before:2020]

Bibliographic Data

Application No: 16/995,607

Foreign Priority claimed: ○ Yes ◉ No

35 USC 119 (a-d) conditions met: ☐ Yes ☑ No ☐ Met After Allowance

Verified and Acknowledged: /Laura B Goddard/ Examiner's Signature | Initials

Title: DETERMINATION OF BASE MODIFICATIONS OF NUCLEIC ACIDS

FILING or 371(c) DATE	CLASS	GROUP ART UNIT	ATTORNEY DOCKET NO.
08/17/2020	435	1642	080015-028410US-1199353
RULE			

APPLICANTS

The Chinese University of Hong Kong, Shatin, HONG KONG

INVENTORS

Yuk-Ming Dennis Lo, Shatin, New Territories, HONG KONG

Rossa Wai Kwun Chiu, Shatin, New Territories, HONG KONG

Kwan Chee Chan, Shatin, New Territories, HONG KONG

Peiyong Jiang, Shatin, New Territories, HONG KONG

Suk Hang Cheng, Shatin, New Territories, HONG KONG

Wenlei Peng, Shatin, New Territories, HONG KONG

On Yee Tse, Fanling, New Territories, Hong Kong SAR, CHINA

CONTINUING DATA

This application has PRO of 63051210 07/13/2020

This application has PRO of 63019790 05/04/2020

This application has PRO of 62991891 03/19/2020

This application has PRO of 62970586 02/05/2020

This application has PRO of 62887987 08/16/2019

FOREIGN APPLICATIONS

IF REQUIRED, FOREIGN LICENSE GRANTED**

09/24/2020

**** SMALL ENTITY ****

STATE OR COUNTRY

HONG KONG

ADDRESS

Kilpatrick Townsend & Stockton LLP - West Coast
Mailstop: IP Docketing - 22
1100 Peachtree Street
Suite 2800
Atlanta, GA 30309
UNITED STATES

FILING FEE RECEIVED

$3,525

WEST Search History

Hide Items | Restore | Clear | Close

DATE: Tuesday, March 23, 2021

Hide?	Set Name	Query	Hit Count
		DB=PGPB,USPT,USOC,EPAB,JPAB,DWPI; PLUR=YES; OP=OR	
☐	L6	L5 AND (pulse interpulse)[CLM]	13
☐	L5	L4 AND @ay<=2020	203
☐	L4	(SMRT "single molecule real time sequencing") AND (methylation methylated) AND nucleotide AND width AND pulse AND (interpulse duration)	203
☐	L3	L1 AND (nucleotide AND width AND pulse AND (interpulse duration))[CLM]	1
☐	L2	L1 AND (C12Q1/6816 AND C12N9/22 AND G16B20/00 AND G16B30/00 AND C12Q1/6851).cpc.	1
☐	L1	(C12Q1/6816 or C12N9/22 or G16B20/00 or G16B30/00 or C12Q1/6851).cpc. AND (SMRT "single molecule real time sequencing") AND (methylation methylated)	444

END OF SEARCH HISTORY

I hereby certify that this correspondence is being filed via EFS-Web with the United States Patent and Trademark Office on March 12, 2021.

KILPATRICK TOWNSEND & STOCKTON LLP

By: /Heather A. Orosco/
Heather A. Orosco

PATENT
Attorney Docket No.: 080015-028410US-1199353
Client Ref. No.: 19/MED/901 US Reg

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

In re Application of:

Yuk-Ming Dennis Lo et al.

Application No.: 16/995,607

Filed: August 17, 2020

For: DETERMINATION OF BASE MODIFICATIONS OF NUCLEIC ACIDS

Customer No.: 20350

Confirmation No.: 1256

Examiner: Laura B. Goddard

Technology Center/Art Unit: 1642

AMENDMENT / RESPONSE TO OFFICE ACTION

Mail Stop Amendment
Commissioner for Patents
P.O. Box 1450
Alexandria, VA 22313-1450

Commissioner:

In response to the Office Action mailed December 30, 2020, please enter the following amendments and consider the following remarks:

Amendments to the Claims are reflected in the listing of claims which begins on page 2 of this paper.

Amendments to the Drawings begin on page 7 of this paper and include an attached Replacement Sheet.

Applicant Summary of Interview with the Examiner begins on page 8 of this paper.

Remarks/Arguments begin on page 9 of this paper.

AMENDMENTS TO THE CLAIMS

This listing of claims will replace all prior versions, and listings of claims in the application.

Listing of Claims:

1. (Currently Amended) A method for detecting a modification of a nucleotide in a nucleic acid molecule, the method comprising:

(a) ~~receiving an input data structure, the input data structure corresponding to a window of nucleotides sequenced in a sample nucleic acid molecule, wherein the sample nucleic acid molecule is sequenced by~~ receiving data acquired by measuring pulses in an optical signal corresponding to [[the]]nucleotides sequenced in a sample nucleic acid molecule~~, the input data structure comprising~~ and obtaining, from the data, values for the following properties:

for each nucleotide~~ within the window~~:

an identity of the nucleotide,

a position of the nucleotide within the sample nucleic acid molecule~~ with respect to a target position within the respective window~~,

a width of the pulse corresponding to the nucleotide, and

an interpulse duration representing a time between the pulse corresponding to the nucleotide and a pulse corresponding to a neighboring nucleotide;

(b) creating an input data structure, the input data structure comprising a window of the nucleotides sequenced in the sample nucleic acid molecule, wherein the input data structure includes, for each nucleotide within the window, the properties:

the identity of the nucleotide,

a position of the nucleotide with respect to a target position within the window,

the width of the pulse corresponding to the nucleotide, and

the interpulse duration;

(c) inputting the input data structure into a model, the model trained by:

receiving a first plurality of first data structures, each first data structure of the first plurality of data structures corresponding to a respective window of nucleotides sequenced in a respective nucleic acid molecule of a plurality of first nucleic acid molecules, wherein each of the first nucleic acid molecules is sequenced by measuring pulses in the optical signal corresponding to the nucleotides, wherein the modification has a known first state in a nucleotide at a target position in each window of each first nucleic acid molecule, each first data structure comprising values for the same properties as the input data structure,

storing a plurality of first training samples, each including one of the first plurality of first data structures and a first label indicating the first state of the nucleotide at the target position, and

optimizing, using the plurality of first training samples, parameters of the model based on outputs of the model matching or not matching corresponding labels of the first labels when the first plurality of first data structures is input to the model, wherein an output of the model specifies whether the nucleotide at the target position in the respective window has the modification,

(d) determining, using the model, whether the modification is present in a nucleotide at the target position within the window in the input data structure.

2. (Original) The method of claim 1, wherein:

the input data structure is one input data structure of a plurality of input data structures,

the sample nucleic acid molecule is one sample nucleic acid molecule of a plurality of sample nucleic acid molecules,

the plurality of sample nucleic acid molecules are obtained from a biological sample of a subject, and

each input data structure corresponds to a respective window of nucleotides sequenced in a respective sample nucleic acid molecule of the plurality of sample nucleic acid molecules, and

the method further comprising:

receiving the plurality of input data structures,

inputting the plurality of input data structures into the model, and

determining, using the model, whether a modification is present in a nucleotide at a target location in the respective window of each input data structure.

3. (Original) The method of claim 2, further comprising:

determining the modification is present at one or more nucleotides, and

determining a classification of a disorder using the presence of the modification at one or more nucleotides.

4. (Original) The method of claim 3, wherein the disorder comprises cancer.

5. (Currently Amended) The method of claim [[3]]4, further comprising:

determining that the classification of the disorder is that the subject has the disorder, and

treating the subject for the disorder by chemotherapy, radiation, or surgery.

6. (Original) The method of claim 3, wherein determining the classification of the disorder uses the number of modifications or the sites of the modifications.

7. (Currently Amended) The method of claim 2, wherein the modification is a methylation,

the method further comprising:

determining the modification is present at one or more nucleotides, and

determining a clinically-relevant DNA fraction, a fetal methylation profile, a maternal methylation profile, a presence of an imprinting gene region, or a tissue of origin using the presence of the modification at one or more nucleotides.

8. (Original) The method of claim 2, wherein each sample nucleic acid molecule of the plurality of sample nucleic acid molecules has a size greater than a cutoff size.

9. (Original) The method of claim 2, wherein:

the plurality of sample nucleic acid molecules align to a plurality of genomic regions,

for each genomic region of the plurality of genomic regions:

a number of sample nucleic acid molecules is aligned to the genomic region,

the number of sample nucleic acid molecules is greater than a cutoff number.

10. (Original) The method of claim 1, further comprising sequencing the sample nucleic acid molecule.

11. (Original) The method of claim 1, wherein the model includes a machine learning model, a principal component analysis, a convolutional neural network, or a logistic regression.

12. (Original) The method of claim 1, wherein:

the window of nucleotides corresponding to the input data structure comprises nucleotides on a first strand of the sample nucleic acid molecule and nucleotides on a second strand of the sample nucleic acid molecule, and

the input data structure further comprises for each nucleotide within the window a value of a strand property, the strand property indicating the nucleotide being present on either the first strand or the second strand.

13. (Original) The method of claim 12, wherein the sample nucleic acid molecule is a circular DNA molecule formed by:

cutting a double-stranded DNA molecule using a Cas9 complex to form a cut double-stranded DNA molecule, and

ligating a hairpin adaptor onto an end of the cut double-stranded DNA molecule.

14. (Original) The method of claim 1, wherein the nucleotides within the window are determined using a circular consensus sequence and without alignment of the sequenced nucleotides to a reference genome.

15. (Original) The method of claim 1, wherein each nucleotide within the window is enriched or filtered.

16. (Original) The method of claim 15, wherein each nucleotide within the window is enriched by:

cutting a double-stranded DNA molecule using a Cas9 complex to form a cut double-stranded DNA molecule, and ligating a hairpin adaptor onto an end of the cut double-stranded DNA molecule, or

filtered by:

selecting double-stranded DNA molecules having a size with a size range.

17. (Original) The method of claim 1, wherein nucleotides within the window are determined without using a circular consensus sequence and without alignment of the sequenced nucleotides to a reference genome.

18-69. (Canceled)

70. (Original) The method of claim 1, wherein the optical signal is a fluorescence signal from a dye-labeled nucleotide.

71. (Original) The method of claim 1, wherein each window associated with the first plurality of data structures comprises 4 consecutive nucleotides on a first strand of each first nucleic acid molecule.

AMENDMENT(S) TO THE DRAWINGS

The attached sheet of drawings includes changes to FIG. 76A. This sheet, which includes FIGS. 76A and B, replaces the original sheet including FIGS. 76A and B. For FIG. 76A, the figure number labeling has been corrected.

Attachment(s): Replacement Sheet

APPLICANT SUMMARY OF INTERVIEW WITH THE EXAMINER

Applicant's representative thanks Examiner Goddard for the telephonic interview on March 4, 2021. During the interview, the parties discussed differences between the Application and the cited references. Aspects of the discussion are incorporated into the claim amendments and remarks below.

REMARKS/ARGUMENTS

Upon entry of the present amendment, claims 1-17, 70, and 71 will be pending in this application. Based on the following remarks, Applicant respectfully requests reconsideration and allowance of the pending claims.

Drawings

The replacement drawing sheet 80 of 145 ("80/145") includes an amendment to FIG. 76A. FIG. 76A has been amended to correct the figure number labeling. Due to the simplicity of the amendment, an annotated sheet showing the changes is not included. No new matter has been added.

Amended Claims

Claim 1 has been amended to clarify elements of the claim. Support can at least be found in the original claim.

Claim 5 has been amended to depend from another claim and to recite particular treatments. Support can be found at least in paragraphs [0524] and [0525].

Claim 7 has been amended to recite that the modification is methylation, which has support throughout the application, including paragraph [0002].

No new matter is added.

Claim Rejections – 35 USC § 112

Claims 3-9 stand rejected under 35 U.S.C. § 112(b) as allegedly being indefinite for failing to particularly point out and distinctly claim the subject matter which the Applicant regards as the invention.

Claims 3-6

With claims 3-6, the Office objects to the phrasing "*using the presence of the modification at one or more nucleotides*." The Office states that the claim does not set forth any

steps of how the modifications are "used," and therefore the claims are unclear what method/process the claim is intended to encompass. This rejection is respectfully traversed.

The Office states that a "claim is indefinite where it merely recites a use without any active, positive steps delimiting how this use is actually practiced." MPEP § 2173.05(q) cites *Ex parte Erlich*, 3 USPQ2d 1011 (Bd. Pat. App & Inter. 1986) for this proposition. *Erlich* involved method claims reciting only a using step and no other steps. In fact, *Erlich* distinguishes a previous appeal, *Ex parte Bull*, 117 USPQ 302 (PTO Bd. App. 1957), where claims were definite because *Bull* also recited positive steps such as "providing" and "maintaining." Similar to claims in *Bull* and different from claims in *Erlich*, present claims 3 and 7 include an active positive step: "*determining the modification is present at one or more nucleotides*." Claims 4-6 depend on claim 3 and some of the claims recite additional active steps. As these claims do not merely recite a use and the claims do recite an active, positive step, the claims are definite.

Furthermore, paragraph [0520] of the Application as filed describes how a classification of a disorder may be determined using the presence of the modification. The description of the various example ways makes clear the genus covered by claim 3, and thus the scope of these claims is clear. Applicant notes that the breadth of a claim does not make the claim indefinite. See MPEP § 2173.04.

Claim 7

The Office objects to claim 7 for allegedly being indefinite as to how "*a clinically-relevant DNA fraction, a fetal methylation profile, a maternal methylation profile, a presence of an imprinting gene region, or a tissue of origin*" can be determined using the presence of the modification at one or more nucleotides. Applicant submits that determining these characteristics are clear in light of the specification.

For example, when the modification is a methylation, determining "*a fetal methylation profile*" and "*a maternal methylation profile*" are clear. For instance, paragraph [0162] describes methylation profiles. A "*clinically-relevant DNA fraction*" includes a fetal DNA fraction, which is correlated with a degree of hypomethylation ([0447]). Regarding determining "*a presence of an imprinting gene region*," DNA methylation is known to be associated with imprinting gene regions (e.g., [0316]-[0317], FIG. 49). Methylation profiles can be used to detect "*a tissue of*

origin" (e.g., [0329], [0338]-[0342], FIG. 58). One of skill in the art would understand how the presence of modifications can be used to determine the characteristics recited in claim 7.

In the interest of expediting prosecution, claim 7 has been amended to recite that the modification is a methylation. Applicants respectfully request withdrawal of this rejection.

Claims 8 and 9

The Office objects to "cutoff" in claims 8 and 9. Claim 8 recites "*a cutoff size*," and claim 9 recites "*a cutoff number*." The Office asserts that the metes and bounds of the claim cannot be determined.

"Cutoff" is explained in [0168]: "a cutoff size can refer to a size above which fragments are excluded"; and "cutoff or threshold may be 'a reference value' or derived from a reference value that is representative of a particular classification or discriminates between two or more classifications." Claim 8 involves a plurality of sample nucleic acid molecules with sizes greater than a cutoff size; nucleic acid molecules with sizes less than the cutoff size are excluded. Claim 9 involves genomic regions having a number of sample nucleic acid molecules greater than a cutoff number; genomic regions having fewer sample nucleic acid molecules are excluded.

The Office appears to reject the claims for not reciting a specific size or number. However, no such recitation is required for definiteness. As stated in paragraphs [0518] and [0519], different cutoff sizes and cutoff numbers can be used. Accordingly, the metes and bounds are clear, and the claims are definite. Hence, claims 8 and 9 are definite.

Thus, Applicant respectfully requests withdrawal of this rejection.

Claim Rejections – 35 USC § 101

Claims 1-17, 70, and 71 stand rejected under 35 U.S.C. § 101 as allegedly being directed to an abstract idea and law of nature without significantly more. Whether a claim is directed to patent eligible subject matter is a two-step inquiry. *Alice Corp. Pty. Ltd. v. CLS Bank Int'l*, 573 U.S. 208, 217-18 (2014).

I. STEP 1 OF ALICE (STEP 2A OF MPEP)

The first step asks whether the claim is directed to a judicial exception ("patent-ineligible concept"). The Office asserts that the claims are directed to an abstract idea.

A. PRONG ONE: EVALUATE WHETHER THE CLAIM RECITES A JUDICIAL EXCEPTION

In Prong One of Step 2A, the Office evaluates whether the claim recites a judicial exception. The Office asserts that elements of the claims are directed to "abstract idea of mathematically manipulating and comparing previously obtained optical pulse sequencing data (for example, obtained by single-molecule, real-time (SMRT) sequencing) and displaying it in a window (presumably on a computer screen)" (Office Action, p. 6). The Office appears to assert that the claims are directed to the "mathematical concepts" grouping of abstract ideas.

Under MPEP § 2106.07(a)(I), a § 101 rejection should point to the specific claim limitation(s) that recite the judicial exception. The Office has quoted the entirety of the claim, with certain portions of the claim bolded. It is not clear what claim limitations are alleged to recite the abstract idea. The alleged abstract idea includes "displaying it in a window," which is **not** recited by the claim. Additionally, the alleged abstract idea includes "comparing previously obtained optical pulse sequencing data." No such limitation in claim 1 recites comparing previously obtained optical pulse sequencing data.

Under MPEP § 2106.04(a), examiners should identify the grouping of abstract ideas. The Office Action appears to identify the abstract idea as "mathematical concepts," given the alleged abstract idea of "mathematically manipulating ... previously obtained optical pulse sequencing data." The Office is not clear as to what mathematical concepts are recited. The bolded limitations of the claim include "*receiving an input data structure*," "*inputting the input data structure into a model*," "*storing a plurality of first training samples*," "*optimizing, using the plurality of first training samples, parameters of the model based on outputs of the model matching or not matching corresponding labels*," and "*determining, using the model, whether the modification is present in a nucleotide at the target position*." However, steps of "*receiving*,"

"*inputting*," "*storing*," and "*determining*" do not recite mathematical concepts. These steps are therefore additional to any alleged judicial exception.

As a result, the Office appears to be interpreting "*optimizing, using the plurality of first training samples, parameters of the model based on outputs of the model matching or not matching corresponding labels*" as directed to a mathematical concept. It is not clear what mathematical concept is recited in this step. The Office may be improperly interpreting any step that may involve math as being directed to a mathematical concept. MPEP § 2106.04(II)(A)(1) is clear that merely involving math is not sufficient for a claim to be directed to a mathematical concept.

In Example 38 of the 2019 USPTO Guidance, "generating a normally distributed first random value" was not considered to be a mathematical concept, but instead is based on a mathematical concept, which is not sufficient. Example 38 states that the example claim did not recite a mathematical relationship, formula, or calculation. In addition, the claim in Example 38 also recites "simulating a first digital representation," which also was not considered to recite a mathematical relationship. Both "generating a normally distributed first random value" and "simulating a first digital representation" involve performing mathematical calculations but performing such calculations are not sufficient for the claim to be directed to a mathematical concept. Example 38 involves using math to obtain a property of simulated circuit. At most, present claim 1 involves math to obtain a property of a nucleotide of a nucleic acid. The involvement of math is not sufficient for a claim to be directed to a mathematical concept.

During the interview, the Examiner noted that mathematical concepts includes "mathematical relationships." MPEP § 2106.04(a)(2)(I)(A) provides a specific definition for "mathematical relationships":

> "A mathematical relationship is a relationship between variables or numbers. A mathematical relationship may be expressed in words or using mathematical symbols. For example, pressure (p) can be described as the ratio between the magnitude of the normal force (F) and area of the surface on contact (A), or it can be set forth in the form of an equation such as p = F/A."

The claim does not recite such a relationship between variables, certainly not similar to the expression of the example of the relationship between force, areas, and pressure.

Accordingly, Applicant submits that the present claims are not directed to an abstract idea.

B. PRONG TWO: EVALUATE WHETHER THE JUDICIAL EXCEPTION IS INTEGRATED INTO A PRACTICAL APPLICATION

Assuming *arguendo* that the claims do recite a judicial exception, Prong Two involves evaluating whether the **claim as a whole** integrates the recited judicial exception into a practical application of the exception.

1. Improvement to technology or technical field

The claims integrate any alleged judicial exception into a practical application because the claims improve biological and medical technologies and/or technical fields and apply the alleged judicial exception in a meaningful way such that the claim as a whole is more than a drafting effort to monopolize the exception.

The present claims improve the technology or technical fields of biology and/or medicine by developing a method to detect a modification of a nucleotide (e.g., a methylation) without certain chemical conversion or nucleic acid amplification (Application, [0006], [0007]). By avoiding some chemical conversion and amplification, certain false negatives, which may occur at a rate of around 3%, for modifications may be avoided ([0180]), resulting in more accurate detection. Methods described herein can achieve 99.3% specificity and 82.6% sensitivity ([0292]) and 90% overall accuracy ([0328]). Additionally, bisulfite sequencing is used to detect methylated cytosines and therefore cannot detect, accurately or otherwise, other types of methylations. The present claims can apply to other types of methylations including methylated adenosines, guanines, and thymines ([0007]), thereby improving accuracy over bisulfite sequencing.

More accurate detection is an improvement to technology. For example, in *CardioNet v. Infobionic*, 2019-1149 (Fed. Cir. 2020), the Federal Circuit held that an improved cardiac monitoring device was not an abstract idea (p. 13). Specifically, the device resulted in "few false negatives and false positives when used to detect atrial fibrillation or atrial flutter" (p. 9). The

device resulted in improving cardiac monitoring technology, which is patent-eligible subject matter (p. 13). Such improved detection is a practical application.

Additionally, avoiding chemical conversion (bisulfite conversion) ([0180]) avoids procedural steps and reagents, thereby making modification detection cheaper and faster. A cheaper, faster, and more accurate detection method is an improvement to technology.

Furthermore, MPEP § 2106.05(a) states that "[c]omponents or methods, such as measurement devices or techniques, that generate new data" may be sufficient to show an improvement to technology. Claim 1 recites a method/technique that generates new data. The new data includes the input data structure with the values for the recited properties. Thus, claim 1, with the generated new data, is an improvement to technology. Additionally, claim 12 recites that the input data structure includes properties of nucleotides on both strands of a nucleic acid molecule, which is also new data and therefore an improvement to technology.

2. Particular treatment or prophylaxis for a disease or medical condition

MPEP § 2106.04(d)(2) states that a particular treatment or prophylaxis for a disease or medical condition is one way to demonstrate that a judicial exception is integrated into a practical application. MPEP § 2106.04(d)(2). Claim 5 recites "*treating the subject for the disorder*." The Office asserts that the claim is "a generic instruction to apply the judicial exception" with a "generic treatment step." In the interest of expediting prosecution, claim 5 specifies that the disorder is cancer and that the treatment is by "chemotherapy, radiation, or surgery," which are particular treatments. As a result, claim 5 integrates the judicial exception into a practical application.

For at least these reasons, the claims are not directed to a judicial exception because any alleged judicial exception is integrated into a practical application. Thus, Applicant respectfully requests withdrawal of this rejection.

II. STEP 2 OF ALICE (STEP 2B OF MPEP)

MPEP § 2106.05(I) directs examiners at Step 2B to evaluate additional elements to determine whether they amount to an inventive concept requires considering them both

individually and in combination to ensure that they amount to significantly more than the judicial exception itself. Additional element(s) or combination of elements that are not well-understood, routine, or conventional activities favors eligibility. MPEP § 2106.05(I).

MPEP § 2106.07(a)(III) makes it clear that a conclusion that an element (or combination of elements) is well-understood, routine, conventional activity must be supported by:

(a) An express statement in the specification, or made by an applicant during prosecution, demonstrating the well-understood, routine, conventional nature of the additional element(s).
(b) One or more of the court decisions discussed in MPEP §2106.05(d)(II) noting the well-understood, routine, conventional nature of the additional element(s).
(c) A publication demonstrating the well-understood, routine, conventional nature of the additional element(s).
(d) Official notice of the well-understood, routine, conventional nature of the additional element(s) (*Not presently alleged.)*

As explained above with Step 2A, Prong One, the Office's arguments regarding the claim reciting a mathematical concept at best applies to the "*optimizing*" step. As a result, "*receiving data acquired by measuring pulses in an optical signal*," "*obtaining values for*" certain properties, "*creating an input data structure*" and "*inputting the input data structure into a model*" are additional limitations to the alleged judicial exception of a mathematical concept. Claim 1 includes the following recitations regarding the data and input data structure:

> *obtaining, from the data, values for the following properties:*
> *for each nucleotide:*
> *an identity of the nucleotide,*
> *a position of the nucleotide within the sample nucleic acid molcule,*
> *a width of the pulse corresponding to the nucleotide, and*
> *an interpulse duration representing a time between the pulse corresponding to the nucleotide and a pulse corresponding to a neighboring nucleotide;*
> *(b) creating an input data structure, the input data structure comprising a window of the nucleotides sequenced in the sample nucleic acid molecule, wherein the input data structure includes, for each nucleotide within the window, the properties:*
> *the identity of the nucleotide,*
> *a position of the nucleotide with respect to a target position within the window,*
> *the width of the pulse corresponding to the nucleotide, and*
> *the interpulse duration.*

The input data structure is not well-understood, routine, or conventional at least because obtaining all the values for the listed properties is not well-understood, routine, or conventional, let alone combining into an input data structure and inputting into a model. As a result, "*receiving data acquired by measuring pulses in an optical signal*," "*obtaining, from the data, values for*" certain properties, "*creating an input data structure*" and "*inputting the input data structure into a model*" are not well-understood, routine, or conventional activity alone or in combination with other limitations of the claim. MPEP § 2106.07(a)(III) makes clear that a conclusion that an element (or combination of elements) is well-understood, routine, conventional activity must be supported by specific types of evidence, none of which is provided in the Office Action. Thus, by including additional elements that are not well-understood, routine, or conventional activity, claim 1 amounts to significantly more than any alleged judicial exception.

For at least these reasons, the claims are directed to eligible subject matter.

Claim Rejections – 35 USC § 102

Claims 1, 10, 12-14, 70, and 71 stand rejected under 35 U.S.C. § 102(a)(1) as allegedly being anticipated by Flusberg et al. (Nature Methods, 2010, 7:461-465) ("Flusberg"). This rejection is respectfully traversed. Flusberg does not disclose "*an identity of the nucleotide*" or "*a width of the pulse corresponding to the nucleotide*" recited in independent claim 1.

Claim 1 recites an input data structure comprising values for "*an identity of the nucleotide*." The Office cites to Figure 1 of Flusberg. Figure 1 of Flusberg illustrates fluorescence traces for samples, which includes merely labeling peaks with the identity of the nucleotide. However, Flusberg does not disclose that the identity of the nucleotide is used in an input data structure to determine whether a modification is present in a nucleotide. Additionally, Flusberg involved DNA templates with the same sequence (p. 462, paragraph bridging col. 1 and 2). Identical DNA templates would not provide information on how different nucleotides affect detection of methylation. Thus, Flusberg does not disclose "*an identity of the nucleotide*" in claim 1.

Additionally, claim 1 recites an input data structure comprising values for "*a width of the pulse corresponding to the nucleotide*." Flusberg does not use the width of the pulse

corresponding to a nucleotide. The sole mention of pulse width in Flusberg is in Supplemental Figure 1. Supplemental Figure 1 plots pulse width ratios for 5-methylcytosine (mC) and 5-hydroxymethylcytosine (hmC) and compares results between the two. However, Flusberg does not use pulse width for determining whether a modification is present in a nucleotide. Thus, Flusberg does not disclose "*a width of the pulse corresponding to the nucleotide*" in claim 1.

For at least these reasons, claim 1 is novel over Flusberg. Claims 10, 12-14, 70, and 71 depend on claim 1 and should be novel as dependent claims. Thus, Applicant respectfully requests withdrawal of this rejection.

Claim Rejection – 35 USC § 103 over Flusberg '338

Claims 1, 2, 8-17, 70, and 71 stand rejected under 35 U.S.C. § 103 as allegedly being obvious over U.S. Patent No. 9,175,338 to Flusberg et al. ("Flusberg '338"). This rejection is respectfully traversed. Flusberg '338 does not teach or suggest "*an identity of the nucleotide*" or "*a window of the nucleotides*" recited in claim 1.

Flusberg '338 does not use an input data structure comprising values for "*an identity of the nucleotide*." The Office Action is not clear what in Flusberg '338 specifically teaches or suggests using the identity of the nucleotide. Flusberg '338 does not teach or suggest that the identity of the nucleotides in a window can be used to determine a modification. For example, in FIGS. 2A and 2B, Flusberg '338 illustrates that a modification may affect the interpulse duration. However, FIG. 2 illustrates no differences between the pulse widths of each of the nucleotides nor any differences in the pulse as a result of the identity of the nucleotide. Hence, one of skill would have no reason to include the identity of nucleotides in a window in an input data structure. The present Application shows that a method that includes the identities of the nucleotides neighboring the methylation site results in significantly improved accuracy ([0259]-[0261], FIG. 29).

In addition, the Office Action is not clear what is the "window" in Flusberg '338. Claim 1 recites "*a window of the nucleotides sequenced in the sample nucleic acid molecule*" and that "*each nucleotide within the window*" provides certain properties. FIGS. 4-8 of the present application show examples of windows. In some instances, the Office Action (p. 13) interprets

window as a display of an output: Flusberg '338 "explicitly suggest their data can be graphically electronically displayed on a computer or printed out." A graphical output is not the window in claim 1.

Additionally, the Office Action cites col. 27, lines 5-23, which states that "an IPD [interpulse duration] averaged over a window of about 20, ..., or 100 bases is used." The use of a window for averaging IPD is not consistent with the window in claim 1, which includes interpulse durations for individual nucleotides (not an average), along with identities, positions, and pulse widths corresponding to individual nucleotides. The other cited mention of window in col. 63, lines 31-50 refers to a window to count incorporated nucleotides and does not disclose the data structure corresponding to the window in claim 1. As Flusberg '338 does not teach the window in claim 21, Flusberg '338 also does not teach an input data structure with "*a position of the nucleotide with respect to a target position within the window*."

Hence, Flusberg '338 does not teach or suggest "*obtaining, from the data, values for the following properties: ... an identity of the nucleotide*," "*creating an input data structure, the input data structure comprising a window of the nucleotides sequenced in the sample nucleic acid molecule*," or "*wherein the input data structure includes, for each nucleotide within the window, ... the identity of the nucleotide, a position of the nucleotide with respect to a target position within the window*."

For at least these reasons, claim 1 is non-obvious over Flusberg '338 as cited. Claims 2, 8-17, 70, and 71 depend on claim 1 and should be non-obvious as dependent claims. Furthermore, at least dependent claims 8, 9, and 12 are non-obvious over Flusberg '338 for additional reasons.

Claim 8

Claim 8 recites that "*each sample nucleic acid molecule of the plurality of sample nucleic acid molecules has a size greater than a cutoff size*." The Office Action cites to portions of Flusberg '338 describing enrichment, but the portions do not disclose a cutoff size for enrichment. The enrichment techniques described in the cited sections involve selecting a particular fragment (e.g., those with a modification). The enrichment techniques do not select for a cutoff size. Accordingly, claim 8 is non-obvious for at least this additional reason.

Claim 9

Claim 9 recites "*the number of sample nucleic acid molecules is greater than a cutoff number*" for a genomic region. The Office Action cited portions of Flusberg '338 regarding assigning a number/barcode to a genomic region. The labeling of genomic regions with a number does not disclose the cutoff number for the number of sample nucleic acid molecules. Accordingly, claim 9 is non-obvious for at least this additional reason.

Claim 12

Claim 12 recites that the input data structure includes information about nucleotides from both strands of a sample nucleic acid molecule. The Office Action cites Flusberg '338 describing circular consensus sequencing. However, these sections of Flusberg '338 do not disclose the specific input data structure recited in claim 12. First, Flusberg does not disclose a time window. And consensus sequencing just uses separate measurements to determine a consensus at a position, as opposed to using a single input data structure for a model to obtain a single measurement. Accordingly, claim 12 is non-obvious for at least this additional reason.

Thus, Applicant respectfully requests withdrawal of this rejection.

Claim Rejection – 35 USC § 103 over Flusberg '338 and Sugai

Claims 3-6 stand rejected under 35 U.S.C. § 103 as allegedly being obvious over Flusberg '338, as applied to claims 1, 2, 8-17, 70, and 71 above, and further in view of Sugai et al. (Clinical Epigenetics, 2017, 9:55, internet pages 1-10) ("Sugai"). Claims 3-6 depend on claim 1, which is non-obvious as explained above. Sugai fails to cure this defect. Sugai is cited for discussion of methylation levels and risks of colorectal cancer and is silent on the difference identified above. For at least these reasons, claims 3-6 are non-obvious over Flusberg '338 and Sugai. Thus, Applicant respectfully requests withdrawal of this rejection.

Claim Rejection – 35 USC § 103 over Flusberg '338 and Kile

Claim 7 stands rejected under 35 U.S.C. § 103 as allegedly being obvious over Flusberg '338, as applied to claims 1, 2, 8-17, 70, and 71 above, and further in view of Kile et al. (PLoS ONE, 2010, 5:e13730, internet pages 1-9) ("Kile"). Claim 7 depends on claim 1, which is non-

obvious, as explained above. Kile fails to cure this defect. Kile is cited for discussion of methylation profiles and is silent on the difference identified above. For at least these reasons, claim 7 is non-obvious over Flusberg '338 and Kile. Thus, Applicant respectfully requests withdrawal of this rejection.

CONCLUSION

In view of the foregoing, Applicant believes all claims now pending in this application are in condition for allowance. The issuance of a formal Notice of Allowance at an early date is respectfully requested.

Except for the issue fees payable under 37 C.F.R. § 1.18, the Director is authorized to charge any additional fees during pendency of this application, including any required extension of time fees, or credit any overpayment to Deposit Account Number 20-1430. This paragraph is intended to be a constructive petition for extension of time in accordance with 37 C.F.R. § 1.136(a)(3).

If the Examiner believes a telephone conference would expedite prosecution of this application, please contact the undersigned at (303) 571-4000 or dhsu@kilpatricktownsend.com.

Respectfully submitted,

/David D. Hsu/
David D. Hsu
Registration No. 71659

KILPATRICK TOWNSEND & STOCKTON LLP

DDH:hao

74401353V.1

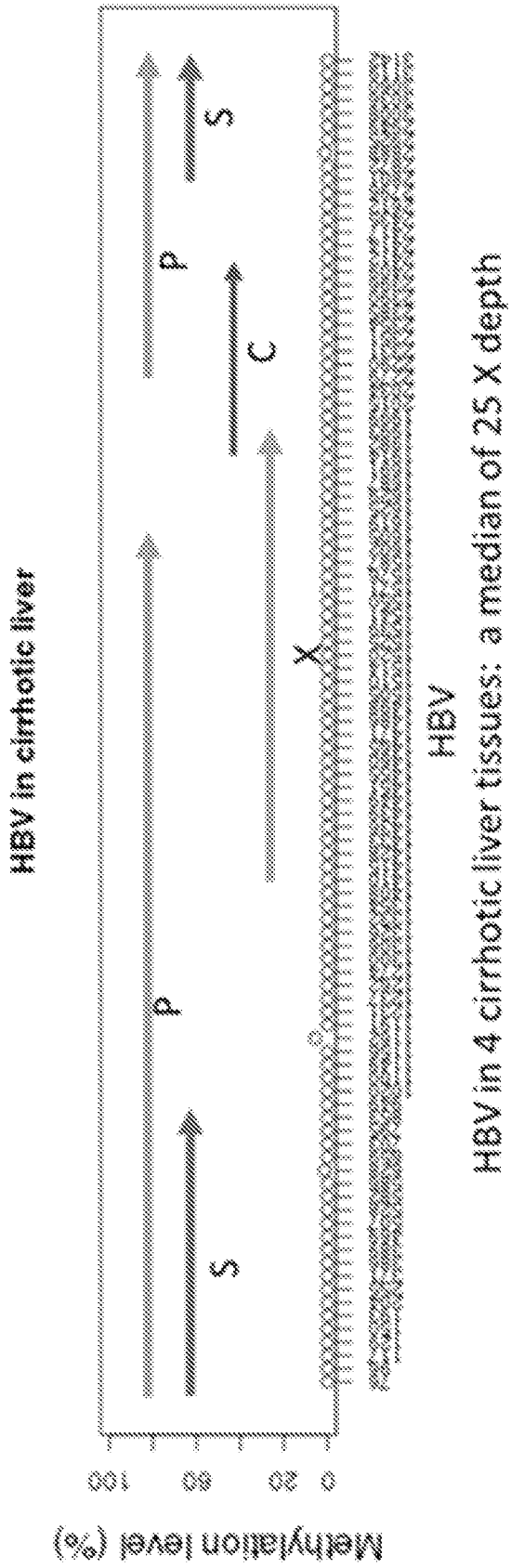


FIG. 76A

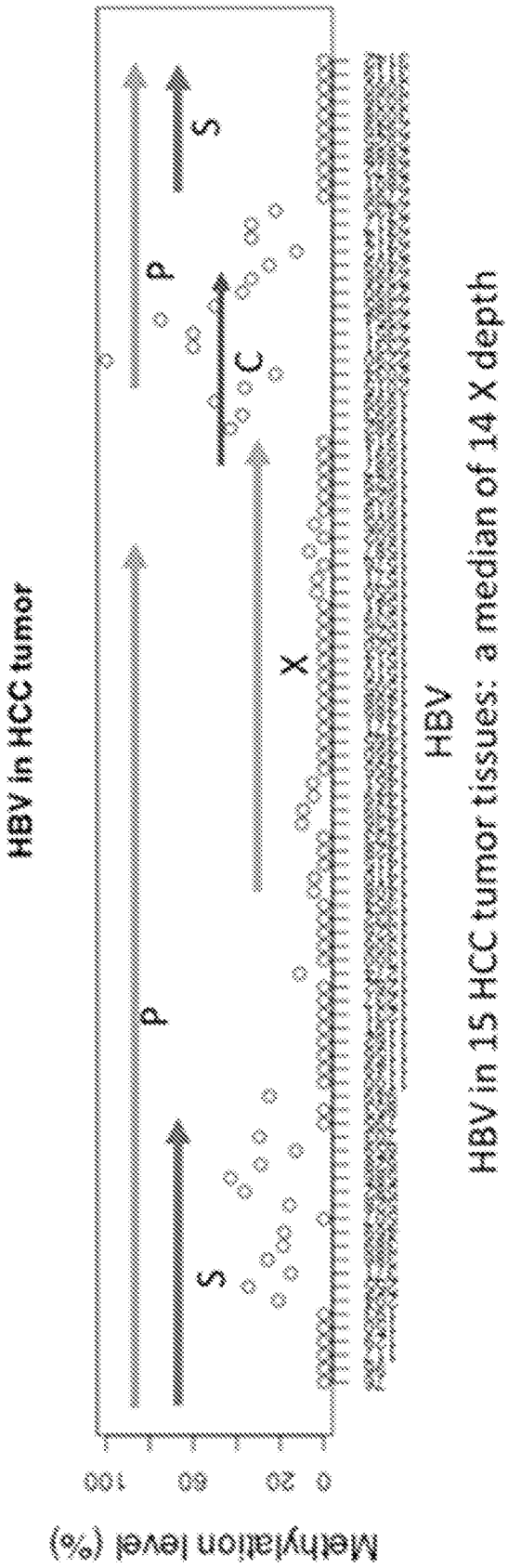


FIG. 76B

Electronic Acknowledgement Receipt

EFS ID:	42171464
Application Number:	16995607
International Application Number:	
Confirmation Number:	1256
Title of Invention:	DETERMINATION OF BASE MODIFICATIONS OF NUCLEIC ACIDS
First Named Inventor/Applicant Name:	Yuk-Ming Dennis Lo
Customer Number:	20350
Filer:	David Dah-Wei Hsu/Heather Orosco
Filer Authorized By:	David Dah-Wei Hsu
Attorney Docket Number:	080015-028410US-1199353
Receipt Date:	12-MAR-2021
Filing Date:	17-AUG-2020
Time Stamp:	17:52:01
Application Type:	Utility under 35 USC 111(a)

Payment information:

Submitted with Payment	no

File Listing:

Document Number	Document Description	File Name	File Size(Bytes)/ Message Digest	Multi Part /.zip	Pages (if appl.)
1		1199353_Response_to_Non-Final_Office_Action.pdf	205025 a2f052a646b03e3644ff5bfa12da41b0dfff8d65	yes	21

	Multipart Description/PDF files in .zip description		
	Document Description	**Start**	**End**
	Amendment/Req. Reconsideration-After Non-Final Reject	1	1
	Claims	2	6
	Drawings-other than black and white line drawings	7	7
	Applicant summary of interview with examiner	8	8
	Applicant Arguments/Remarks Made in an Amendment	9	21

Warnings:					
Information:					
2	Drawings-other than black and white line drawings	1199353_Replacement_Sheet_for_FIG_76A.pdf	383440 d903821175459f9c38107704fa9855e197ade4b7	no	1
Warnings:					
Information:					
		Total Files Size (in bytes):	588465		

This Acknowledgement Receipt evidences receipt on the noted date by the USPTO of the indicated documents, characterized by the applicant, and including page counts, where applicable. It serves as evidence of receipt similar to a Post Card, as described in MPEP 503.

New Applications Under 35 U.S.C. 111

If a new application is being filed and the application includes the necessary components for a filing date (see 37 CFR 1.53(b)-(d) and MPEP 506), a Filing Receipt (37 CFR 1.54) will be issued in due course and the date shown on this Acknowledgement Receipt will establish the filing date of the application.

National Stage of an International Application under 35 U.S.C. 371

If a timely submission to enter the national stage of an international application is compliant with the conditions of 35 U.S.C. 371 and other applicable requirements a Form PCT/DO/EO/903 indicating acceptance of the application as a national stage submission under 35 U.S.C. 371 will be issued in addition to the Filing Receipt, in due course.

New International Application Filed with the USPTO as a Receiving Office

If a new international application is being filed and the international application includes the necessary components for an international filing date (see PCT Article 11 and MPEP 1810), a Notification of the International Application Number and of the International Filing Date (Form PCT/RO/105) will be issued in due course, subject to prescriptions concerning national security, and the date shown on this Acknowledgement Receipt will establish the international filing date of the application.

SCORE Placeholder Sheet for IFW Content

Application Number: 16995607 Document Date: 03/12/2021

The presence of this form in the IFW record indicates that the following document type was received in electronic format on the date identified above. This content is stored in the SCORE database.

Since this was an electronic submission, there is no physical artifact folder, no artifact folder is recorded in PALM, and no paper documents or physical media exist. The TIFF images in the IFW record were created from the original documents that are stored in SCORE.

- Drawing

At the time of document entry (noted above):

- USPTO employees may access SCORE content via DAV or via the SCORE web page.
- External customers may access SCORE content via PAIR using the Supplemental Content tab.

Form Revision Date: March 1, 2019

PTO/SB/06 (09-11)
Approved for use through 1/31/2014. OMB 0651-0032
U.S. Patent and Trademark Office; U.S. DEPARTMENT OF COMMERCE
Under the Paperwork Reduction Act of 1995, no persons are required to respond to a collection of information unless it displays a valid OMB control number.

PATENT APPLICATION FEE DETERMINATION RECORD Substitute for Form PTO-875	Application or Docket Number 16/995,607	Filing Date 08/17/2020	☐ To be Mailed

ENTITY: ☐ LARGE ☑ SMALL ☐ MICRO

APPLICATION AS FILED - PART I

FOR	(Column 1) NUMBER FILED	(Column 2) NUMBER EXTRA		RATE ($)	FEE ($)
☐ BASIC FEE (37 CFR 1.16(a), (b), or (c))	N/A	N/A		N/A	
☐ SEARCH FEE (37 CFR 1.16(k), (i), or (m))	N/A	N/A		N/A	
☐ EXAMINATION FEE (37 CFR 1.16(o), (p), or (q))	N/A	N/A		N/A	
TOTAL CLAIMS (37 CFR 1.16(i))	minus 20 =	*		x $50 =	
INDEPENDENT CLAIMS (37 CFR 1.16(h))	minus 3 =	*		x $230 =	
☐ APPLICATION SIZE FEE (37 CFR 1.16(s))	If the specification and drawings exceed 100 sheets of paper, the application size fee due is $310 ($155 for small entity) for each additional 50 sheets or fraction thereof. See 35 U.S.C. 41(a)(1)(G) and 37 CFR 1.16(s).				
☐ MULTIPLE DEPENDENT CLAIM PRESENT (37 CFR 1.16(j))					
* If the difference in column 1 is less than zero, enter "0" in column 2.				TOTAL	

APPLICATION AS AMENDED - PART II

		(Column 1)		(Column 2)	(Column 3)			
AMENDMENT	03/12/2021	CLAIMS REMAINING AFTER AMENDMENT		HIGHEST NUMBER PREVIOUSLY PAID FOR	PRESENT EXTRA		RATE ($)	ADDITIONAL FEE ($)
	Total (37 CFR 1.16(i))	* 19	Minus	** 20	= 0		x $50 =	0
	Independent (37 CFR 1.16(h))	* 1	Minus	*** 3	= 0		x $240 =	0
	☐ Application Size Fee (37 CFR 1.16(s))							
	☐ FIRST PRESENTATION OF MULTIPLE DEPENDENT CLAIM (37 CFR 1.16(j))							
							TOTAL ADD'L FEE	0

		(Column 1)		(Column 2)	(Column 3)			
AMENDMENT		CLAIMS REMAINING AFTER AMENDMENT		HIGHEST NUMBER PREVIOUSLY PAID FOR	PRESENT EXTRA		RATE ($)	ADDITIONAL FEE ($)
	Total (37 CFR 1.16(i))	*	Minus	**	=		x $0 =	
	Independent (37 CFR 1.16(h))	*	Minus	***	=		x $0 =	
	☐ Application Size Fee (37 CFR 1.16(s))							
	☐ FIRST PRESENTATION OF MULTIPLE DEPENDENT CLAIM (37 CFR 1.16(j))							
							TOTAL ADD'L FEE	

* If the entry in column 1 is less than the entry in column 2, write "0" in column 3.	LIE
** If the "Highest Number Previously Paid For" IN THIS SPACE is less than 20, enter "20".	/MARSHA R RICHARDS/
*** If the "Highest Number Previously Paid For" IN THIS SPACE is less than 3, enter "3".	
The "Highest Number Previously Paid For" (Total or Independent) is the highest number found in the appropriate box in column 1.	

This collection of information is required by 37 CFR 1.16. The information is required to obtain or retain a benefit by the public which is to file (and by the USPTO to process) an application. Confidentiality is governed by 35 U.S.C. 122 and 37 CFR 1.14. This collection is estimated to take 12 minutes to complete, including gathering, preparing, and submitting the completed application form to the USPTO. Time will vary depending upon the individual case. Any comments on the amount of time you require to complete this form and/or suggestions for reducing this burden, should be sent to the Chief Information Officer, U.S. Patent and Trademark Office, U.S. Department of Commerce, P.O. Box 1450, Alexandria, VA 22313-1450. DO NOT SEND FEES OR COMPLETED FORMS TO THIS ADDRESS. **SEND TO: Commissioner for Patents, P.O. Box 1450, Alexandria, VA 22313-1450.**

If you need assistance in completing the form, call 1-800-PTO-9199 and select option 2.

UNITED STATES PATENT AND TRADEMARK OFFICE

UNITED STATES DEPARTMENT OF COMMERCE
United States Patent and Trademark Office
Address: COMMISSIONER FOR PATENTS
P.O. Box 1450
Alexandria, Virginia 22313-1450
www.uspto.gov

APPLICATION NO.	FILING DATE	FIRST NAMED INVENTOR	ATTORNEY DOCKET NO.	CONFIRMATION NO.
16/995,607	08/17/2020	Yuk-Ming Dennis Lo	080015-028410US-1199353	1256

20350 7590 03/09/2021

Kilpatrick Townsend & Stockton LLP - West Coast
Mailstop: IP Docketing - 22
1100 Peachtree Street
Suite 2800
Atlanta, GA 30309

EXAMINER	
GODDARD, LAURA B	

ART UNIT	PAPER NUMBER
1642	

NOTIFICATION DATE	DELIVERY MODE
03/09/2021	ELECTRONIC

Please find below and/or attached an Office communication concerning this application or proceeding.

The time period for reply, if any, is set in the attached communication.

Notice of the Office communication was sent electronically on above-indicated "Notification Date" to the following e-mail address(es):

KTSDocketing2@kilpatrick.foundationip.com
ipefiling@kilpatricktownsend.com

Applicant-Initiated Interview Summary	Application No. 16/995,607	Applicant(s) Lo et al.		
	Examiner LAURA B GODDARD	Art Unit 1642	AIA (First Inventor to File) Status Yes	Page 1 of 2

All Participants (applicant, applicants representative, PTO personnel)	Title	Type
LAURA B GODDARD	Primary Examiner	Telephonic
Mr. David Hsu	Attorney of Record	

Date of Interview: 04 March 2021

Issues Discussed:

35 U.S.C. 101

Examiner suggested an amendment to include active steps of: (a) sequencing the sample of nucleic acid molecule by measuring pulses in an optical signal corresponding to the nucleotides in the sample and obtaining the properties for each nucleotide in the sample: an identity of the nucleotide, a numerical position of the nucleotide, a width of the pulse corresponding to the nucleotide; and an interpulse duration representing the time between the pulse corresponding to the nucleotide and a pulse corresponding to a neighboring nucleotide; (b) creating an input data structure, the input data structure comprising a window of the nucleotides sequenced in the sample nucleic acid molecule, wherein for each nucleotide within the window providing the properties obtained in step (a); (c) inputting the input data structure into a model...(d) determining using the model, whether the modification is present ... Applicants did not want an active step of sequencing but suggested an active step of measuring the properties from previously obtained sequencing data for input into the model. Examiner believes measuring and utilizing the width of the pulse in determining the presence of mutations is a step in the right direction of providing a novel element or step in the method. Examiner will consult a 101 expert when the amendments are submitted.

35 U.S.C. 112

Examiner will consider the amendments when entered.

35 U.S.C. 102

Examiner stated the method of Flusberg displays various widths of pulses in their sequence data in a window, however it does not appear that they measure the widths as a data point to be combined with interpulse duration, nucleotide identity and position in determining the presence of mutations. Examiner will consider and compare future claim amendments with the prior art.

35 U.S.C. 103

See comments in 102 above.

Applicant-Initiated Interview Summary	**Application No.** 16/995,607	**Applicant(s)** Lo et al.		
	Examiner LAURA B GODDARD	**Art Unit** 1642	**AIA (First Inventor to File) Status** Yes	**Page** **2 of 2**

/Laura B Goddard/ Primary Examiner, Art Unit 1642	

Applicant is reminded that a complete written statement as to the substance of the interview must be made of record in the application file. It is the applicants responsibility to provide the written statement, unless the interview was initiated by the Examiner and the Examiner has indicated that a written summary will be provided. See MPEP 713.04
Please further see:
MPEP 713.04
Title 37 Code of Federal Regulations (CFR) § 1.133 Interviews, paragraph (b)
37 CFR § 1.2 Business to be transacted in writing

Applicant recordation instructions: The formal written reply to the last Office action must include the substance of the interview. (See MPEP section 713.04). If a reply to the last Office action has already been filed, applicant is given a non-extendable period of the longer of one month or thirty days from this interview date, or the mailing date of this interview summary form, whichever is later, to file a statement of the substance of the interview.

Examiner recordation instructions: Examiners must summarize the substance of any interview of record. A complete and proper recordation of the substance of an interview should include the items listed in MPEP 713.04 for complete and proper recordation including the identification of the general thrust of each argument or issue discussed, a general indication of any other pertinent matters discussed regarding patentability and the general results or outcome of the interview, to include an indication as to whether or not agreement was reached on the issues raised.

UNITED STATES PATENT AND TRADEMARK OFFICE

USPTO Automated Interview Request (AIR)

Feb 23 2021

This paper requesting to schedule and/or conduct an interview is appropriate because:

This submission is requested to be accepted as an authorization for this interview to communicate via the internet. Recognizing that Internet communications are not secure, I hereby authorize the USPTO to communicate with the undersigned concerning scheduling of the interview via video conference, instant messaging, or electronic mail, and to conduct the interview in accordance with office practice including video conferencing.

Name(s):
David Hsu

S-signature:
/David Hsu/

Registration Number:
71659

U.S. Application Number:
16995607

Confirmation Number:
1256

E-mail Address:
dhsu@kilpatricktownsend.com

Phone Number:
+1 3036073298

Proposed Time of Interview:
3-3-2021 11:00 AM ET

Alternative Proposed Time(s) of Interview:
3-3-2021 3:00 PM ET

Alternative Proposed Time(s) of Interview:
3-4-2021 11:00 AM ET

Prefered Interview Type:
Telephonic

I am the applicant or applicant's representative for this application.

Topic for Discussion:
112, 102, 103, 101 rejections (detailed agenda to be sent prior to interview)

UNITED STATES PATENT AND TRADEMARK OFFICE

UNITED STATES DEPARTMENT OF COMMERCE
United States Patent and Trademark Office
Address: COMMISSIONER FOR PATENTS
P.O. Box 1450
Alexandria, Virginia 22313-1450
www.uspto.gov

APPLICATION NO.	FILING DATE	FIRST NAMED INVENTOR	ATTORNEY DOCKET NO.	CONFIRMATION NO.
16/995,607	08/17/2020	Yuk-Ming Dennis Lo	080015-028410US-1199353	1256

20350 7590 02/22/2021

Kilpatrick Townsend & Stockton LLP - West Coast
Mailstop: IP Docketing - 22
1100 Peachtree Street
Suite 2800
Atlanta, GA 30309

EXAMINER
GODDARD, LAURA B

ART UNIT	PAPER NUMBER
1642	

NOTIFICATION DATE	DELIVERY MODE
02/22/2021	ELECTRONIC

Please find below and/or attached an Office communication concerning this application or proceeding.

The time period for reply, if any, is set in the attached communication.

Notice of the Office communication was sent electronically on above-indicated "Notification Date" to the following e-mail address(es):

KTSDocketing2@kilpatrick.foundationip.com
ipefiling@kilpatricktownsend.com

UNITED STATES PATENT AND TRADEMARK OFFICE

Commissioner for Patents
United States Patent and Trademark Office
P.O. Box 1450
Alexandria, VA 22313-1450
www.uspto.gov

Kilpatrick Townsend & Stockton LLP - West Coast Mailstop: IP Docketing - 22
1100 Peachtree Street Suite 2800
Atlanta, GA 30309

In re Application of	:
Lo et al.	: **DECISION ON PETITION**
Application No. 16/995,607	: **UNDER 37 C.F.R. § 1.84(a)(2)**
Filed: 08/17/2020	:
Attorney Docket No. 080015-028410US-1199353	:

This is a decision on the Petition to Accept Color Drawings under 37 C.F.R 1.84 (a)(2), received in the United States Patent and Trademark Office (USPTO) on August 17, 2020.

A grantable petition under 37 C.F.R. 1.84(a)(2) must meet the following requirements:

1. Payment of the fee set forth under 37 C.F.R. 1.17(h),
2. Submission of three (3) sets of color drawings, or one (1) set of color drawings if filed by EFS-Web, and
3. The first paragraph of the brief description of the drawings in the Specification contains the required language pertaining to colored drawings.

A review of U.S. Patent and Trademark records indicates that each of the above three requirements have been satisfied. Accordingly, the petition under 37 C.F.R. 1.84(a)(2) is **GRANTED**.

Should there be any questions about this decision, please contact MARIANNE C SEIDEL, by letter addressed to Director, TC 1600, at the address listed above, or by telephone at (571)272-0584, or by facsimile sent to the general Office facsimile number, 571-273-8300.

/MARIANNE C SEIDEL/
Quality Assurance Sp

/Zachariah Lucas/
Supervisory Patent Examiner, Art Unit 1600

Technology Center Petition Count Sheet

Application No.	Applicant(s)	
16/995,607	Lo et al.	

Petition Code	Petition Type
☐ 601	Petition to Make Special, Age or Health (37 CFR 1.102, MPEP § 708.02)
☐ 604	Petition to Make Special, Accelerated Examination (37 CFR 1.102, MPEP § 708.02)
☐ 607	Petition to Make Special, For Reasons not Provided in 601-606 and 639-640 (37 CFR 1.102, MPEP § 708.02)
☐ 608	Petition to Reopen After a Board Decision (37 CFR 1.198)
☐ 609	Petition for Review of Final Restriction Requirement (37 CFR 1.144)
☐ 610	Petition Invoking Authority of Commissioner under 37 CFR 1.181 Not Specified in Codes 611-617
☐ 611	Petition Relating to Prematureness of Final Rejection (37 CFR 1.181, MPEP § 706.07)
☐ 612	Petition Relating to Refusal to Enter an Amendment (37 CFR 1.181 and 1.127, MPEP § 714.09)
☐ 613	Petition Relating to Withdraw of holding of Abandonment (37 CFR 1.137, MPEP § 711.03)
☐ 614	Petition Relating to Withdraw of holding of Abandonment (37 CFR 1.137, MPEP § 7.14.19)
☐ 615	Petition Relating to Formal Sufficiency / Propriety of Affidavits (37 CFR 1.131, 1.132, 1.608, MPEP § 608.04(C)
☐ 617	Petition Relating to Refusal to Enter an Amendment (37 CFR 1.312)
☐ 620	Petition for Extension of Time (37 CFR 1.136 (b))
☐ 622	Petition Concerning Appeal Application before Transfer of Jurisdiction to Board of Appeals (MPEP § 1206)
☐ 623	Petition for Second or Subsequent Suspension of Action (37 CFR 1.103, MPEP § 709)
☐ 634	Petition to Correct Inventorship in a Patent not in Interference (37 CFR 1.324, MPEP § 1481)
☐ 635	Petition to Correct Inventorship in a Patent Application (37 CFR 1.48)
☐ 636	Petition to Change Inventorship in a Patent (37 CFR 1.324)
☐ 637	Petition to Withdraw from Issue, before Payment of an Issue Fee (37 CFR 1.313(a))
☐ 639	Petition for Make Special Prospective Manufacture (37 CFR 1.102, MPEP § 708.02)
☐ 646	Petition for Make Special Accelerated Examination
☑ 690	Petition Related Color Drawings of Photographs
☐ 699	Petition for Matters Before Technology Center

Information Required for OPESS Processing

☑ GRANTED ☐ GRANTED-IN-PART ☐ DENIED ☐ DISMISSED

☑ Mail Decision and Enter Decision in Petition Database/PALM
☐ Copy Decision, Send to Scanning
☐ Fees Need to be Charged
☐ _____

INFORMATION DISCLOSURE STATEMENT BY APPLICANT (Not for submission under 37 CFR 1.99)		
	Application Number	16/995,607
	Filing Date	August 17, 2020
	First Named Inventor	Yuk-Ming Dennis Lo
	Art Unit	1642
	Examiner Name	Laura B. Goddard
	Attorney Docket Number	080015-028410US-1199353

U.S. PATENTS						
Examiner Initial*	Cite No	Patent Number	Kind Code [1]	Issue Date	Name of Patentee or Applicant of cited Document	Pages, Columns, Lines, Where Relevant Passages or Relevant Figures Appear

U.S. PATENT APPLICATION PUBLICATIONS						
Examiner Initial*	Cite No	Publication Number	Kind Code [1]	Publication Date	Name of Patentee or Applicant of cited Document	Pages, Columns, Lines, Where Relevant Passages or Relevant Figures Appear
	1.	2013/0230909	A1	09-05-2013	Pan et al.	

FOREIGN PATENT DOCUMENTS								
Examiner Initial*	Cite No	Foreign Document Number[3]	Country Code[2]	Kind Code[4]	Publication Date	Name of Patentee or Applicant of cited Document	Pages, Columns, Lines where Relevant Passages or Relevant Figures Appear	T[5]
	2.	2010068289	WO	A2	06-17-2010	Pacific Biosciences of California, Inc.		
	3.	2010027484	WO	A2	03-11-2010	Pacific Biosciences of California, Inc.		
	4.	2014153757	WO	A1	10-02-2014	BGI Shenzhen	English machine translation attached.	
	5.	2013188846	WO	A1	12-19-2013	Harry Stylli		
	6.	104053789	CN	A	09-17-2014	BGI Health Service Co., Ltd.	English abstract attached. Equivalent to AU 2012380221 B2 below.	
	7.	104053789	CN	B	02-10-2016	BGI Health Service Co., Ltd.	English abstract attached. Equivalent to AU 2012380221 B2 below.	
	8.	2013170429	WO	A1	11-21-2013	BGI Health Service Co., Ltd.	English abstract attached. Equivalent to AU 2012380221 B2 below.	

INFORMATION DISCLOSURE STATEMENT BY APPLICANT (Not for submission under 37 CFR 1.99)		
	Application Number	16/995,607
	Filing Date	August 17, 2020
	First Named Inventor	Yuk-Ming Dennis Lo
	Art Unit	1642
	Examiner Name	Laura B. Goddard
	Attorney Docket Number	080015-028410US-1199353

	9.	2012380221	AU	B2	09-29-2016	BGI Diagnosis Co., Ltd.	Equivalent to CN 104053789 A & B and to WO 2013170429 A1 above.	

NON-PATENT LITERATURE DOCUMENTS			
Examiner Initials*	Cite No	Include name of the author (in CAPITAL LETTERS), title of the article (when appropriate), title of the item (book, magazine, journal, serial, symposium, catalog, etc.), date, page(s), volume-issue number(s), publisher, city and/or country where published.	T[5]
	10.	International Search Report and Written Opinion mailed November 19, 2020 in International Patent Application No. PCT/CN2020/109602. 11 pages.	

EXAMINER SIGNATURE			
Examiner Signature		Date Considered	

*EXAMINER: Initial if reference considered, whether or not citation is in conformance with MPEP 609. Draw line through a citation if not in conformance and not considered. Include copy of this form with next communication to applicant.

[1] See Kind Codes of USPTO Patent Documents at www.USPTO.GOV or MPEP 901.04. [2] Enter office that issued the document, by the two-letter code (WIPO Standard ST.3). [3] For Japanese patent documents, the indication of the year of the reign of the Emperor must precede the serial number of the patent document.
[4] Kind of document by the appropriate symbols as indicated on the document under WIPO Standard ST.16 if possible. [5] Applicant is to place a check mark here if English language translation is attached.

I hereby certify that this correspondence is being filed via EFS-Web with the United States Patent and Trademark Office on February 19, 2021.

KILPATRICK TOWNSEND & STOCKTON LLP

By: /Heather A. Orosco/
Heather A. Orosco

PATENT
Attorney Docket No.: 080015-028410US-1199353
Client Ref. No.: 19/MED/901 US Reg

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

In re Application of:	Confirmation No.: 1256
Yuk-Ming Dennis Lo	Examiner: Laura B. Goddard
Application No.: 16/995,607	Technology Center/Art Unit: 1642
Filed: August 17, 2020	
For: DETERMINATION OF BASE MODIFICATIONS OF NUCLEIC ACIDS	SUPPLEMENTAL INFORMATION DISCLOSURE STATEMENT
Customer No.: 20350	

Commissioner for Patents
P.O. Box 1450
Alexandria, VA 22313-1450

Commissioner:

The references cited on attached form PTO/SB/08A are being called to the attention of the Examiner. In accordance with the provisions of 37 CFR §1.98(a)(2), copies of any cited U.S. Patent and U.S. Patent Application Publications are not provided. Copies of the remaining cited references are provided.

It is respectfully requested that the cited references be expressly considered during the prosecution of this application, and the references be made of record therein and appear among the "references cited" on any patent to issue therefrom.

As provided for by 37 CFR §1.97(g) and (h), no inference should be made that the information and references cited are prior art merely because they are in this statement and no

representation is being made that a search has been conducted or that this statement encompasses all the possible relevant information.

This IDS is being filed after the mailing of the first Office Action on the merits, but before the mailing date of the final Office Action, Notice of Allowance, or any other action that closes prosecution.

I hereby certify under 37 CFR §1.97(e)(1) that each item of information contained in this Information Disclosure Statement was first cited in any communication from a foreign patent office in a counterpart foreign application not more than three months prior to the filing of this statement.

Applicant believes that no fee is required for submission of this statement. However, if any additional fees are due for the submission of this Information Disclosure Statement, please deduct those fees from Deposit Account No. 20-1430.

Respectfully submitted,

/David D. Hsu/

David D. Hsu
Registration No. 71659

KILPATRICK TOWNSEND & STOCKTON LLP

Attachment

74512556V.1

(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
17 June 2010 (17.06.2010)

PCT

(10) International Publication Number
WO 2010/068289 A2

(51) International Patent Classification:
C12Q 1/68 (2006.01) *C12N 15/10* (2006.01)

(21) International Application Number:
PCT/US2009/006517

(22) International Filing Date:
10 December 2009 (10.12.2009)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
61/201,551 11 December 2008 (11.12.2008) US
61/180,350 21 May 2009 (21.05.2009) US
61/186,661 12 June 2009 (12.06.2009) US

(71) Applicant *(for all designated States except US)*: **PACIFIC BIOSCIENCES OF CALIFORNIA, INC.** [US/US]; 1505 Adams Drive, Menlo Park, CA 94025 (US).

(72) Inventors; and

(75) Inventors/Applicants *(for US only)*: **FLUSBERG, Benjamin** [US/US]; 533 Thain Way, Palo Alto, CA 94306 (US). **TURNER, Stephen** [US/US]; 356 Stanford Avenue, Menlo Park, CA 94025 (US). **LEE, Jessica** [US/US]; 10856 Linda Vista Drive, Cupertino, CA 95014 (US). **JIA, Lei** [CN/CN]; 3415 Janice Way, Palo Alto, CA 94303 (US). **KORLACH, Jonas** [DE/US]; 36100 Toulouse Street, Newark, CA 94560 (US). **SORENSON, Jon** [US/US]; 1725 Nason Street, Alameda, CA 94501 (US). **WEBSTER, Dale** [US/US]; 2301 Ticonderoga Drive, San Mateo, CA 94402 (US). **LYLE, John** [US/US]; 3 Avocet Drive #203, Redwood Shores, CA 94065 (US). **TRAVERS, Kevin** [US/US]; 3385 Gonzaga Place, Menlo Park, CA 95051 (US). **HANES, Jeremiah** [US/US]; 3243 Spring Street, Redwood City, CA 94063 (US).

(74) Agents: **QUINE, Jonathan, Alan** et al.; Quine Intellectual Property Law Group. P.C., P.O. Box 458, Alameda, CA 94501 (US).

(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IS, JP, KE, KG, KM, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PE, PG, PH, PL, PT, RO, RS, RU, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(84) Designated States *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, GH, GM, KE, LS, MW, MZ, NA, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Published:

— *without international search report and to be republished upon receipt of that report (Rule 48.2(g))*

(54) Title: CLASSIFICATION OF NUCLEIC ACID TEMPLATES

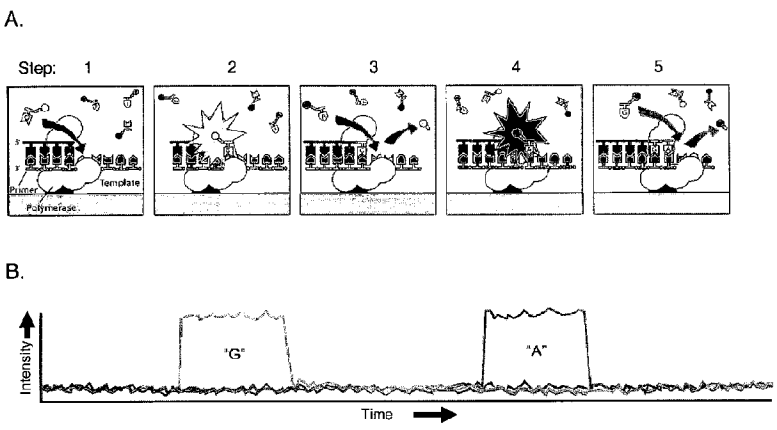


Figure 1

(57) Abstract: Methods, compositions, and systems are provided for characterization of modified nucleic acids. In certain preferred embodiments, single molecule sequencing methods are provided for identification of modified nucleotides within nucleic acid sequences. Modifications detectable by the methods provided herein include chemically modified bases, enzymatically modified bases, abasic sites, non-natural bases, secondary structures, and agents bound to a template nucleic acid.

WO 2010/068289 A2

CLASSIFICATION OF NUCLEIC ACID TEMPLATES

CROSS-REFERENCE TO RELATED APPLICATIONS

[0001] This application claims the benefit of Provisional U.S. Patent Application No. 61/201,551, filed December 11, 2008; Provisional U.S. Patent Application No. 61/180,350, filed May 21, 2009; and Provisional U.S. Patent Application No. 61/186,661, filed June 12, 2009, the full disclosures of which are hereby incorporated herein by reference in their entireties for all purposes.

BACKGROUND OF THE INVENTION

[0002] Assays for analysis of biological processes are exploited for a variety of desired applications. For example, monitoring the activity of key biological pathways can lead to a better understanding of the functioning of those systems as well as those factors that might disrupt the proper functioning of those systems. In fact, various different disease states caused by operation or disruption of specific biological pathways are the focus of much medical research. By understanding these pathways, one can model approaches for affecting them to prevent the onset of the disease or mitigate its effects once manifested.

A stereotypical example of the exploitation of biological process monitoring is in the area of pharmaceutical research and development. In particular, therapeutically relevant biological pathways, or individual steps or subsets of individual steps in those pathways, are often reproduced or modeled in *in vitro* systems to facilitate analysis. By observing the progress of these steps or whole pathways in the presence and absence of potential therapeutic compositions, e.g., pharmaceutical compounds or other materials, one can identify the ability of those compositions to affect the *in vitro* system, and potentially beneficially affect an organism in which the pathway is functioning in a detrimental way. By way of specific example, reversible methylation of the 5' position of cytosine by methyltransferases is one of the most widely studied epigenetic modifications. In mammals, 5-methylcytosine (5-MeC) frequently occurs at CpG dinucleotides, which often cluster in regions called CpG islands that are at or near transcription start sites. Methylation of cytosine in CpG islands can interfere with transcription factor binding and is associated with transcription repression and gene regulation. In addition, DNA methylation is known to be essential for mammalian development and has been associated with cancer and other disease processes. Recently, a new 5-hydroxymethylcytosine epigenetic marker has been identified in certain cell types in the brain, suggesting that it plays a role in

epigenetic control of neuronal function (S. Kriaucionis, et al., *Science* **2009**, 324(5929): 929-30, incorporated herein by reference in its entirety for all purposes). Further information on cytosine methylation and its impact on gene regulation, development, and disease processes is provided in the art, e.g., in A. Bird, *Genes Dev* **2002**, *16*, 6; M. Gardiner-Garden, et al., *J Mol Biol* **1987**, *196*, 261; S. Saxonov, et al., *Proc Natl Acad Sci U S A* **2006**, *103*, 1412; R. Jaenisch, et al., *Nat Genet* **2003**, *33 Suppl*, 245; E. Li, et al., *Cell* **1992**, *69*, 915; A. Razin, et al., *Hum Mol Genet* **1995**, *4 Spec No*, 1751; P. A. Jones, et al., *Nat Rev Genet* **2002**, *3*, 415; P. A. Jones, et al., *Nat Genet* **1999**, *21*, 163; and K. D. Robertson, *Nat Rev Genet* **2005**, *6*, 597, all of which are incorporated herein by reference in their entireties for all purposes.

[0004] In contrast to determining a human genome, mapping of the human methylome is a more complex task because the methylation status differs between tissue types, changes with age, and is altered by environmental factors (P. A. Jones, et al., *Cancer Res* **2005**, *65*, 11241, incorporated herein by reference in its entirety for all purposes). Comprehensive, high-resolution determination of genome-wide methylation patterns from a given sample has been challenging due to the sample preparation demands and short read lengths characteristic of current DNA sequencing technologies (K. R. Pomraning, et al., *Methods* **2009**, *47*, 142, incorporated herein by reference in its entirety for all purposes).

[0005] Bisulfite sequencing is the current method of choice for single-nucleotide resolution methylation profiling (S. Beck, et al., *Trends Genet* **2008**, *24*, 231; and S. J. Cokus, et al., *Nature* **2008**, 452, 215, the disclosures of which are incorporated herein by reference in their entireties for all purposes). Treatment of DNA with bisulfite converts unmethylated cytosine, but not 5-MeC, to uracil (M. Frommer, et al., *Proc Natl Acad Sci U S A* **1992**, *89*, 1827, incorporated herein by reference in its entirety for all purposes). The DNA is then amplified (which converts all uracils into thymines) and subsequently analyzed with various methods, including microarray-based techniques (R. S. Gitan, et al., *Genome Res* **2002**, *12*, 158, incorporated herein by reference in its entirety for all purposes) or 2nd-generation sequencing (K. H. Taylor, et al., *Cancer Res* **2007**, *67*, 8511 ; and R. Lister, et al., *Cell* **2008**, *133*, 523, both incorporated herein by reference in their entireties for all purposes). While bisulfite-based techniques have greatly advanced the analysis of methylated DNA, they also have several drawbacks. First, bisulfite sequencing requires a significant amount of sample preparation time (K. R. Pomraning, et al., *supra*). Second, the harsh reaction conditions necessary for complete conversion of unmethylated cytosine to uracil lead to degradation of DNA (C. Grunau, et al., *Nucleic Acids Res* **2001**, *29*, E65, incorporated herein by reference in its entirety for all purposes), and thus necessitate large starting amounts of the sample, which can be problematic for some applications.

[0006] Furthermore, because bisulfite sequencing relies on either microarray or 2nd-generation DNA sequencing technologies for its readout of methylation status, it also suffers from the same limitations as do these methodologies. For array-based procedures, the reduction in sequence complexity caused by bisulfite conversion makes it difficult to design enough unique probes for genome-wide profiling (S. Beck, et al., *supra*). Most 2nd-generation DNA sequencing techniques employ short reads and thus have difficulties aligning to highly repetitive genomic regions (K. R. Pomraning, et al., *supra*). This is especially problematic, since many CpG islands reside in such regions. Given these limitations, bisulfite sequencing is also not well suited for *de novo* methylation profiling (S. Beck, et al., *supra*).

[0007] In another widely used technique, methylated DNA immunoprecipitation (MeDIP), an antibody against 5-MeC is used to enrich for methylated DNA sequences (M. Weber, et al., *Nat Genet* **2005**, *37*, 853, incorporated herein by reference in its entirety for all purposes). MeDIP has many advantageous attributes for genome-wide assessment of methylation status, but it does not offer as high base resolution as bisulfite treatment-based methods. In addition, it is also hampered by the same limitations of current microarray and 2nd-generation sequencing technologies.

[0008] Research efforts aimed at increasing our understanding of the human methylome would benefit greatly from the development of a new methylation profiling technology that does not suffer from the limitations described above. Accordingly, there exists a need for improved techniques for detection of modifications in nucleic acid sequences, and particularly nucleic acid methylation.

[0009] Typically, modeled biological systems rely on bulk reactions that ascertain general trends of biological reactions and provide indications of how such bulk systems react to different effectors. While such systems are useful as models of bulk reactions *in vivo*, a substantial amount of information is lost in the averaging of these bulk reaction results. In particular, the activity of and effects on individual molecular complexes cannot generally be teased out of such bulk data collection strategies.

[0010] Single-molecule real-time analysis of nucleic acid synthesis has been shown to provide powerful advantages over nucleic acid synthesis monitoring that is commonly exploited in sequencing processes. In particular, by concurrently monitoring the synthesis process of nucleic acid polymerases as they work in replicating nucleic acids, one gains advantages of a system that has been perfected over millions of years of evolution. In particular, the natural DNA synthesis processes provide the ability to replicate whole genomes in extremely short periods of time, and do so with an extremely high level of fidelity to the underlying template being replicated.

[0011] The present invention is directed to a variety of different single-molecule real-time analyses for monitoring the progress and effectors of biological reactions, and in particular detecting modifications in nucleic acid sequences. For example, the present invention provides a direct methylation sequencing technology based on observing the kinetics of single polymerase molecules in real time and with high multiplex. This technique will provide for fast and economical analysis of methylation patterns, even in repetitive genomic regions.

BRIEF SUMMARY OF THE INVENTION

[0012] The present invention is generally directed to the detection of modified nucleic acid sequences, and particularly the detection of methylated bases within nucleic acid sequences using a real time direct detection of such methylated sites. The present invention is expected to have a major impact on research aiming to illuminate the role of DNA methylation in human health.

[0013] In certain aspects of the invention, methods are provided for identification of a modification in a nucleic acid molecule. In general, a template nucleic acid comprising the modification and an enzyme capable of processing the template are provided. The template nucleic acid is contacted with the enzyme, and the subsequent processing of the template by the enzyme is monitored. A change in the processing is detected, and this change is indicative of the presence of the modification in the template. Exemplary modifications that can be detected by the methods of the invention include, but are not limited to methylated bases (e.g., 5-methylcytosine, N6-methyladenosine, etc.), pseudouridine bases, 7,8-dihydro-8-oxoguanine bases, 2'-O-methyl derivative bases, nicks, apurinic sites, apyrimidic sites, pyrimidine dimers, a cis-platen crosslinking products, oxidation damage, hydrolysis damage, bulky base adducts, thymine dimers, photochemistry reaction products, interstrand crosslinking products, mismatched bases, secondary structures, and bound agents. In preferred embodiments, nucleotides or analogs thereof that are incorporated into a nascent strand synthesized by the enzyme are distinctly labeled to allow identification of a sequence of specific nucleotides or nucleotide analogs so incorporated. In certain preferred embodiments, labels are linked to nucleotides or nucleotide analogs through a phosphate group, e.g., a phosphate group other than the alpha phosphate group. As such, the labels are removed from the nucleotide or nucleotide analog upon incorporation into the nascent strand.

[0014] In some embodiments, the template nucleic acid is treated prior to processing by the enzyme, e.g., to alter the modification. The treatment may be chemical or enzymatic, and includes, e.g., glycosylase modification, bisulfite modification, DMS modification, cytosine methyltransferase modification, TET1 modification, and cytidine deaminase modification. In

some embodiments, non-natural nucleotide analogs (e.g., pyrene analogs) are incorporated into a nascent strand synthesized by the enzyme. In some embodiments, the methods comprise both treatment of the template and incorporation of non-natural nucleotide analogs into the nascent strand. In some embodiments, non-natural nucleotides are incorporated into a nascent strand in a position to pair with a modification in the template. For example, a methylated cytosine in the template can be paired with a modified guanine nucleotide analog; a template modification can pair with a non-natural nucleotide analog to form a non-natural base pair, e.g., isocytosine and isoguanine; 5-methylisocytosine and isoguanine; Im-N^O and Im-O^N; A* and T*; and 8-oxoG and adenine. In some embodiments, non-incorporatable nucleotide analogs bind the template/enzyme complex, but are not incorporated into the nascent strand, and detection of this "nonproductive" binding serves as an indication of the modification in the template. Such non-incorporatable nucleotide analogs are preferably distinctly labeled to facilitate monitoring, and optionally to distinguish such binding from incorporation of incorporatable nucleotide analogs that comprise labels.

[0015] In certain embodiments, the template nucleic acid comprises regions of internal complementarity (e.g., a double-stranded portion) and at least one single-stranded portion, and preferably the modification is located within at least one of the regions of internal complementarity. In certain embodiments, the template is a circular template. In certain embodiments, the template is a circular template comprising at least two regions of internal complementarity. In certain embodiments, the enzyme is a polymerase, such as a DNA polymerase, and RNA polymerase, a reverse transcriptase, or a derivative or variant thereof. In preferred embodiments, the enzyme is a polymerase enzyme capable of strand displacement. In specific embodiments, the enzyme is a Φ29 polymerase, optionally comprising at least one mutation at a position selected from the group consisting of K392, K422, I93, M188, K392, V399, T421, K422; S95, Y101, M102; Q99, L123, K124, T189, A190; G191, S388; P127, L384, N387, S388; and L389, Y390, and G391.

[0016] Examples of changes in the processing of the template by the enzyme that are monitored in various embodiments of the invention include, but are not limited to, kinetics, processivity, signal characteristics, error metrics, signal context, and the like. In some embodiments, a change occurs only at the modification, and in other embodiments the change occurs at one or more positions proximal to the modification, which may also include the modification position.

[0017] In certain aspects, the methods further comprise mapping the modification. In certain preferred embodiments, mapping the modification comprises analyzing a portion of the sequence read that was generated immediately prior to, during, and/or immediately after detecting the change in processing to determine a sequence complementary to the template

nucleic acid; determining the complement of the sequence complementary to the template nucleic acid; and mapping the modification at a position in the template nucleic acid that is proximal to the complement of the sequence complementary to the template nucleic acid.

[0018] In certain embodiments, a change in the processing that is indicative of the modification is a kinetic difference in the processing (e.g., detected as an alteration in one or more of interpulse duration, interpulse width, processivity, etc.) and/or a change in an error metric (e.g., accuracy, an increase in binding events that do not result in incorporation, etc.) The change in processing can be indicative of the type of modification is present in the template nucleic acid, since different types of modifications have different effects on the activity and/or fidelity of the enzyme.

[0019] In preferred embodiments, the monitoring occurs in real time during the processing of the template by the enzyme. In preferred embodiments, the template nucleic acid and the enzyme form a complex that is immobilized at a reaction site on a substrate, and in more preferred embodiments a plurality of complexes are immobilized at optically resolvable reaction sites on the substrate, wherein a single complex immobilized at one of the reaction sites is optically resolvable from any other of the complexes immobilized at any other of the reaction sites. In certain embodiments, the optically resolvable reaction sites are nanometer-scale apertures in the substrate, and can be optical confinements, such as zero-mode waveguides. In preferred embodiments, the template nucleic acid is plurality of template nucleic acids that are optically resolvable from one another during the monitoring. Preferably, the template nucleic acid is not amplified prior to contacting it with the enzyme.

[0020] In some embodiments, the modification is secondary structure in the template nucleic acid, e.g., a hairpin loop, and the change in the modification is a kinetic change, e.g., an increased interpulse duration or increased pulse width. Certain methods for identifying such a secondary structure generally comprise generating a sequence read for the template nucleic acid before and after the pause; identifying a first portion of the sequence read generated before the pause that is complementary to a second portion of the sequence read generated after the pause; and determining a likelihood that a hairpin loop formed by annealing of the first portion to the second portion was present in the template nucleic acid during the processing based at least upon the nucleotide composition of the first portion and the second portion.

[0021] In another aspect of the invention, methods for detecting binding of an agent to a single nucleic acid template are provided. In certain embodiments, such methods generally comprise providing the single nucleic acid template in complex with a polymerase; introducing a reaction mixture to the complex, wherein the reaction mixture comprises the agent; and monitoring synthesis of a polynucleotide by the polymerase, wherein the polynucleotide is complementary

to the single nucleic acid template, and wherein a change in the synthesis is indicative of binding of the agent to the single nucleic acid template. Examples of agents appropriate for use in such methods include, but are not limited to, transcription factors, polymerases, reverse transcriptases, histones, restriction enzymes, antibodies, nucleic acid binding proteins, and nucleic acid binding agents. Examples of single nucleic acid templates appropriate for use in such methods include, but are not limited to, double-stranded DNA, double-stranded RNA, single-stranded DNA, single-stranded RNA, DNA/RNA hybrids, and templates comprising both double-stranded and single-stranded regions.

[0022] In certain aspects of the invention, a consensus binding site of the agent is determined. This determination can comprise, e.g., performing a plurality of sequencing-by-synthesis reactions on a set of single nucleic acid templates in the presence of the agent to generate a set of binding-affected nascent polynucleotide sequences; performing a plurality of sequencing-by-synthesis reactions on the set of single nucleic acid templates in the absence of the agent to generate a set of full-length nascent polynucleotide sequences; analyzing the binding-affected nascent polynucleotide sequences to determine a location at which the agent bound the single nucleic acid template during the sequencing-by-synthesis reactions in the presence of the agent; and identifying a sequence common to the full-length nascent polynucleotide sequences at the location, thereby identifying the consensus binding site of the agent. In certain embodiments, the binding-affected nascent polynucleotide sequences are truncated nascent polynucleotide sequences; and in other embodiments, the binding-affected nascent polynucleotide sequences are nascent polynucleotide sequences whose synthesis was paused at the location at which the agent bound.

[0023] In yet further aspects of the invention, methods for detecting modifications in a single nucleic acid template during a sequencing-by-synthesis reaction are provided. For example, such a method can comprise providing the single nucleic acid template in complex with a polymerase; introducing a reaction mixture to the complex, wherein the reaction mixture comprises an agent that specifically binds to the modification; and monitoring synthesis of a polynucleotide by the polymerase, wherein the polynucleotide is complementary to the single nucleic acid template, and wherein a pause or cessation of the synthesis of the polynucleotide is indicative of binding of the agent to the single nucleic acid template, thereby detecting the modification in the single nucleic acid template. In certain embodiments, the modification is an 8-oxoG lesion and/or the agent is a protein is selected from the group consisting of hOGG1, FPG, yOGG1, AlkA, Nth, Nei, MutY, UDG, SMUG, TDG, NEIL, an antibody against 8-oxoG, or a binding domain thereof. In other embodiments, the modification is a methylated base and/or the agent is a protein selected from the group consisting of MECP2, MBD1, MBD2, MBD4,

UHRF1, an antibody against the methylated base, or a binding domain thereof. In further embodiments, the modification is a secondary structure formation in the nucleic acid template. Preferably, the complex is immobilized in an optical confinement. The template can comprise, e.g., single-stranded linear nucleic acid, single-stranded circular nucleic acid, double-stranded linear nucleic acid, double-stranded circular nucleic acid, or a combination thereof.

[0024] In certain embodiments, a modification in a template nucleic acid can be repaired by including components of damage repair machinery in the reaction mixture, e.g., during a sequencing-by-synthesis reaction. In certain embodiments, the readlength of the sequencing-by-synthesis reaction is longer than that for a further sequencing-by-synthesis reaction performed with the single nucleic acid template in complex with the polymerase, but absent the agent and the damage repair machinery.

[0025] In other aspects of the invention, methods for bypassing one or more modifications in a single nucleic acid template during a sequencing-by-synthesis reaction are provided. Certain exemplary methods include providing the single nucleic acid template in complex with a sequencing engine; introducing a reaction mixture to the complex, wherein the reaction mixture comprises a bypass polymerase; initiating the sequencing-by-synthesis reaction; monitoring synthesis of a polynucleotide by the sequencing engine, wherein the polynucleotide is complementary to the single nucleic acid template, and wherein a pause or cessation of the synthesis of the polynucleotide is indicative that the sequencing engine has encountered a modification in the single nucleic acid template; subsequently monitoring synthesis of the polynucleotide by the bypass polymerase, which is indicative that the modification is being bypassed; and repeating the monitoring steps each time a further modification is encountered in the single nucleic acid template, thereby bypassing one or more modifications in a single nucleic acid template during a sequencing-by-synthesis reaction. In certain embodiments, the bypass polymerase comprises a detectable label and detection of a signal from the detectable label during the sequencing-by-synthesis reaction is indicative that the bypass polymerase is actively synthesizing the polynucleotide. In preferred embodiments, the readlength of the sequencing-by-synthesis reaction is longer than that for a further sequencing-by-synthesis reaction performed with the single nucleic acid template in complex with the sequencing engine, but absent the bypass polymerase. In specific embodiments, the reaction mixture comprises multiple different bypass polymerases and a processivity factor. Preferably, at least one of the single nucleic acid template, the sequencing engine, and the bypass polymerase is immobilized, directly or indirectly, in an optical confinement. For example, the template can be immobilized by hybridization to an oligonucleotide primer immobilized in the optical confinement. In certain preferred embodiments, the single nucleic acid template is processed by the sequencing engine

multiple times at a single reaction site, and further wherein redundant sequence data is generated. Nucleic acid templates for use with the methods can be circular and/or can comprise multiple copies of a nucleic acid segment of interest. Further, in certain embodiments the sequencing-by-synthesis reaction generates a polynucleotide comprising multiple copies of a segment complementary to the segment of interest, and further wherein redundant sequence data is generated.

[0026] In further aspects, novel compositions are provided. For example, in certain embodiments a composition of the invention comprises a substrate having a reaction site that is optically resolvable from any other reaction site on the substrate; a single complex of a template and sequencing engine immobilized at the reaction site; a mixture of incorporatable nucleotides or nucleotide analogs; and at least one modification in the template nucleic acid, wherein the template at or proximal to the modification is processed differently than the template distal from the modification. In some embodiments the modification is a non-natural base in the template. The modification can be located in either a strand of the template nucleic acid that is complementary to a nascent strand synthesized by the sequencing engine, or a strand of the template nucleic acid that is displaced by the sequencing engine. In certain preferred embodiments, the template nucleic acid comprises internally complementary regions, and optionally, the modification is located within one of the internally complementary regions. Certain embodiments further comprise at least one type of non-incorporatable nucleotide analog. Certain embodiments comprise at least one type of non-natural incorporatable nucleotide analog. Preferably, one or more or all of the nucleotides or nucleotide analogs in a composition of the invention are tagged with distinct labels that distinguish different types of nucleotides or nucleotide analogs from one another. Compositions of the invention can also include an agent other than the sequencing engine that binds to the modification and/or chemically or enzymatically alters the modification. Preferably, compositions of the invention comprise a nascent strand generated by the sequencing engine, wherein the nascent strand is complementary to the template nucleic acid, and, optionally, comprises multiple copies of regions complementary to the template nucleic acid. Further, certain compositions comprise a nanometer-scale aperture in the substrate, where the reaction site is disposed within the nanometer-scale aperture, e.g., a zero-mode waveguide.

[0027] In further aspects of the invention, systems for identification of modifications within a nucleic acid template are provided. In certain preferred embodiments, a system of the invention comprises a solid support having a polymerase complex disposed thereon (e.g., at a reaction site, e.g., in a nanoscale aperture, e.g., in a zero-mode waveguide), the polymerase complex comprising a nucleic acid template comprising a modification; a mounting stage configured to

receive the solid support; an optical train positioned to be in optical communication with at least a portion of the solid support to detect signals emanating therefrom; a translation system operably coupled to the mounting stage or the optical train for moving one of the optical train and the solid support relative to the other; and a data processing system operably coupled to the optical train. Preferably, the polymerase complex comprises a polymerase enzyme that is actively processing the nucleic acid template. More preferably, the polymerase complex comprises a polymerase enzyme that is processively synthesizing a nascent strand by template-directed synthesis. In preferred embodiments, the optical train detects signals emanating from the solid support during the processing of the nucleic acid template.

[0028] In yet further aspects, the invention provides machine-implemented methods for transforming reaction data into modification detection data, wherein the reaction data is representative of a series of events during a sequencing-by-synthesis reaction wherein a nascent strand is synthesized based upon a nucleotide sequence of a template nucleic acid, and the modification detection data is representative of a presence of one or more modifications within a template nucleic acid. Preferably, one or more steps of the machine-implemented method are performed via a user interface implemented in a machine that comprises instructions stored in machine-readable medium and a processor that executes the instructions. In a final aspect of the invention, a computer program products are provided. In certain embodiments, machine-implemented methods for transforming reaction data comprise a classifier to distinguish between true incorporations and stochastic pulses, a segmenting algorithm based on a hidden Markov model architecture, and/or a segmenting algorithm based on a conditional random field framework. In certain specific embodiments, the methods identify regions in the template having a higher density of stochastic pulses than true incorporations. In certain specific embodiments, the methods identify regions in the template having higher IPD. Exemplary computer program products of the invention typically comprise a computer usable medium having a computer readable program code embodied therein, said computer readable program code adapted to be executed to implement the machine-implemented methods of the invention; and the machine-readable medium on which the results of one or more steps of the machine-implemented method are stored.

BRIEF DESCRIPTION OF THE DRAWINGS

[0029] Figure 1 provides an exemplary illustration of single-molecule, real-time (SMRT™) nucleic acid sequencing.

[0030] Figure 2 provides illustrative examples of various types of reaction data in the context of a pulse trace.

[0031] Figure 3 schematically illustrates a structural model of 5-MeC positioned one base in the 5' direction relative to the DNA polymerase active site.

[0032] Figure 4 illustrates an exemplary embodiment of five-base DNA methylation sequencing. Figure 4A depicts fragmentation of genomic DNA to generate the DNA template. Figure 4B illustrates DNA glycosylase excising a 5-MeC from the template.

[0033] Figure 5 provides an illustrative embodiment of a reaction comprising a linear template and a damage-binding agent that recognizes a lesion in a single-stranded template.

[0034] Figure 6 illustrates an embodiment of the invention comprising a circular template and a damage-binding agent that recognizes a lesion in a double-stranded template.

[0035] Figure 7 illustrates an observation of true incorporations (solid line) versus stochastic pulses (dashed line) across time.

[0036] Figure 8 provides an illustrative example of a simple hidden Markov model for classifying pause (P) versus sequencing (S) states within a sequencing trace.

[0037] Figure 9 provides an illustrative example of a system of the invention.

[0038] Figure 10A provides a schematic for exemplary template nucleic acids of the invention. Figure 10B provides graphs plotting interpulse duration for template nucleic acids as depicted in 10A.

[0039] Figure 11 provides a graph plotting IPD ratio against template position for a template nucleic acid comprising 5-methylcytosine modifications.

[0040] Figure 12 provides a graph plotting IPD ratio against template position for a template nucleic acid comprising 5-methylcytosine modifications.

[0041] Figure 13A provides a schematic for exemplary template nucleic acids of the invention. Figure 13B provides graphs plotting interpulse duration for template nucleic acids as depicted in 13A. Figure 13C provides a ROC curve for the data provided in 13B.

[0042] Figure 14 provides a graph plotting IPD ratio against template position for a template nucleic acid comprising N6-methyladenosine modifications.

[0043] Figure 15 provides a graph plotting IPD ratio against template position for a template nucleic acid comprising 5-hydroxymethylcytosine modifications.

[0044] Figure 16 provides a graph plotting pulse width ratio against template position for a template nucleic acid comprising 5-hydroxymethylcytosine modifications.

[0045] Figure 17 provides a graph plotting IPD ratio against template position for a template nucleic acid comprising 8-oxoguanosine modifications.

[0046] Figure 18 provides a graph plotting pulse width ratio against template position for a template nucleic acid comprising 8-oxoguanosine modifications.

DETAILED DESCRIPTION OF THE INVENTION

I. General

[0047] The present invention is generally directed to methods, compositions, and systems for detecting modifications within nucleic acid sequences, and in particularly preferred aspects, methylated nucleotides within sequence templates through the use of single molecule nucleic acid analysis. The ability to detect modifications within nucleic acid sequences is useful for mapping such modifications in various types and/or sets of nucleic acid sequences, e.g., across a set of mRNA transcripts, across a chromosomal region of interest, or across an entire genome. The modifications so mapped can then be related to transcriptional activity, secondary structure of the nucleic acid, siRNA activity, mRNA translation dynamics, kinetics and/or affinities of DNA- and RNA-binding proteins, and other aspects of nucleic acid (e.g., DNA and/or RNA) metabolism.

[0048] Although certain embodiments of the invention are described in terms of detection of modified nucleotides or other modifications in a single-stranded DNA molecule (e.g., a single-stranded template DNA), various aspects of the invention are applicable to many different types of nucleic acids, including e.g., single- and double-stranded nucleic acids that may comprise DNA, RNA (e.g., mRNA, siRNA, microRNA, rRNA, tRNA, snRNA, etc.), RNA-DNA hybrids, PNA, LNA, morpholino, and other RNA and/or DNA mimetics and derivatives thereof, and combinations of any of the foregoing. Nucleic acids for use with the methods, compositions, and systems provided herein may consist entirely of native nucleotides, or may comprise non-natural bases/nucleotides (e.g., synthetic and/or engineered) that may be paired with native nucleotides or may be paired with the same or a different non-natural base/nucleotide. In certain preferred embodiments, the nucleic acid comprises a combination of single-stranded and double-stranded regions, e.g., such as the templates described in U.S.S.N. 12/383,855 and 12/413,258, both filed on March 27, 2009 and incorporated herein by reference in their entireties for all purposes. In particular, mRNA modifications are difficult to detect by technologies that require reverse transcriptase PCR amplification because such treatment does not maintain the modification in the amplicons. The present invention provides methods for analyzing modifications in RNA molecules that do not require such amplification.

[0049] Generally speaking, the methods of the invention involve monitoring of an analytical reaction to collect "reaction data," wherein the reaction data is indicative of the progress of the reaction. Reaction data included data collected directly from the reaction, as well as the results of various manipulations of that directly collected data, any or a combination of which can serve as a signal for the presence of a modification in the template nucleic acid. For example, certain

types of reaction data are collected in real time during the course of the reaction, such as metrics related to reaction kinetics, processivity, signal characteristics, and the like. Signal characteristics vary depending on the type of analytical reaction being monitored. For example, some reactions use detectable labels to tag one or more reaction components, and signal characteristics for a detectable label include, but are not limited to, the type of signal (e.g., wavelength, charge, etc.) and the shape of the signal (e.g., height, width, curve, etc.). Further, signal characteristics for multiple signals (e.g., temporally adjacent signals) can also be used, including, e.g., the distance between signals during a reaction, the number of extra signals (e.g., that do not correspond to the progress of the reaction), internal complementarity, and the local signal context (i.e., one or more signal that precede and/or follow a given signal). For example, template-directed sequencing reactions often combine signal data from multiple nucleotide incorporation events to generate a sequence read for a nascent strand synthesized, and this sequence read is used to derive, e.g., by complementarity, the sequence of the template strand. Other types of reaction data are generated from statistical analysis of real time reaction data, including, e.g., accuracy, precision, conformance, etc. In some embodiments, data from a source other than the reaction being monitored is also used. For example, a sequence read generated during a nucleic acid sequencing reaction can be compared to sequence reads generated in replicate experiments, or to known or derived reference sequences from the same or a related biological source. Alternatively or additionally, a portion of a template nucleic acid preparation can be amplified using unmodified nucleotides and subsequently sequenced to provide an experimental reference sequence to be compared to the sequence of the original template in the absence of amplification. Although certain specific embodiments of the use of particular types of reaction data to detect certain kinds of modifications are described at length herein, it is to be understood that the methods, compositions, and systems are not limited to these specific embodiments. Different types of reaction data can be combined to detected various kinds of modifications, and in certain embodiments more than one type of modification can be detected and identified during a single reaction on a single template. Such variations to the detailed embodiments of the invention will be clear to one of ordinary skill based upon the teachings provided herein.

[0050] In certain embodiments, redundant sequence information is generated and analyzed to detect one or more modifications in a template nucleic acid. Redundancy can be achieved in various ways, including carrying out multiple sequencing reactions using the same original template, e.g., in an array format, e.g., a ZMW array. In some embodiments in which a lesion is unlikely to occur in all the copies of a given template, reaction data (e.g., sequence reads, kinetics, signal characteristics, signal context, and/or results from further statistical analyses)

generated for the multiple reactions can be combined and subjected to statistical analysis to determine a consensus sequence for the template. In this way, the reaction data from a region in a first copy of the template can be supplemented and/or corrected with reaction data from the same region in a second copy of the template. Similarly, a template can be amplified (e.g., via rolling circle amplification) to generate a concatemer comprising multiple copies of the template, and the concatemer can be subjected to sequencing, thereby generating a sequencing read that is internally redundant. As such, the sequence data from a first segment of the concatemer (corresponding to a first region of the template) can be supplemented and/or corrected with sequence data from a second segment of the concatemer also corresponding to the first region of the template. Alternatively or additionally, a template can be subjected to repeated sequencing reactions to generate redundant sequence information that can be analyzed to more thoroughly characterize the modification(s) present in the template.

[0051] The term "modification" as used herein is intended to refer not only to a chemical modification of a nucleic acids, but also to a variation in nucleic acid conformation, interaction of an agent with a nucleic acid (e.g., bound to the nucleic acid), and other perturbations associated with the nucleic acid. As such, a location or position of a modification is a locus (e.g., a single nucleotide or multiple contiguous or noncontiguous nucleotides) at which such modification occurs within the nucleic acid. For a double-stranded template, such a modification may occur in the strand complementary to a nascent strand synthesized by a polymerase processing the template, or may occur in the displaced strand. Although certain specific embodiments of the invention are described in terms of 5-methylcytosine detection, detection of other types of modified nucleotides (e.g., N^6-methyladenosine, N^3-methyladenosine, N^7-methylguanosine, 5-hydroxymethylcytosine, other methylated nucleotides, pseudouridine, thiouridine, isoguanosine, isocytosine, dihydrouridine, queuosine, wyosine, inosine, triazole, diaminopurine, 8-oxoguanosine, and 2'-O-methyl derivatives of adenosine, cytidine, guanosine, and uridine) is also contemplated. These and other modifications are known to those of ordinary skill in the art and are further described, e.g., in Narayan P, et al. (1987) Mol Cell Biol 7(4):1572-5; Horowitz S, et al. (1984) Proc Natl Acad Sci U.S.A. 81(18):5667-71; "RNA's Outfits: The nucleic acid has dozens of chemical costumes," (2009) C&EN; 87(36):65-68; Kriaucionis, et al. (2009) Science 324 (5929): 929–30; and Tahiliani, et al. (2009) Science 324 (5929): 930–35; Matray, et al. (1999) Nature 399(6737):704-8; Ooi, et al. (2008) Cell 133: 1145-8; Petersson, et al. (2005) J Am Chem Soc. 127(5):1424-30; Johnson, et al. (2004) 32(6):1937-41; Kimoto, et al. (2007) Nucleic Acids Res. 35(16):5360-9; Ahle, et al. (2005) Nucleic Acids Res 33(10):3176; Krueger, et al., Curr Opinions in Chem Biology 2007, 11(6):588); Krueger, et al. (2009) Chemistry & Biology 16(3):242; McCullough, et al. (1999)

Annual Rev of Biochem 68:255; and Liu, et al. (2003) Science 302(5646):868-71, the disclosures of which are incorporated herein by reference in their entireties for all purposes. Modifications further include the presence of non-natural base pairs in the template nucleic acid, including but not limited to hydroxypyridone and pyridopurine homo- and hetero-base pairs, pyridine-2,6-dicarboxylate and pyridine metallo-base pairs, pyridine-2,6-dicarboxamide and a pyridine metallo-base pairs, metal-mediated pyrimidine base pairs T-Hg(II)-T and C-Ag(I)-C, and metallo-homo-basepairs of 2,6-bis(ethylthiomethyl)pyridine nucleobases Spy, and alkyne-, enamine-, alcohol-, imidazole-, guanidine-, and pyridyl-substitutions to the purine or pyridimine base (Wettig, et al. (2003) J Inorg Biochem 94:94-99; Clever, et al. (2005) Angew Chem Int Ed 117:7370-7374; Schlegel, et al. (2009) Org Biomol Chem 7(3):476-82; Zimmerman, et al. (2004) Bioorg Chem 32(1):13-25; Yanagida, et al. (2007) Nucleic Acids Symp Ser (Oxf) 51:179-80; Zimmerman (2002) J Am Chem Soc 124(46):13684-5; Buncel, et al. (1985) Inorg Biochem 25:61-73; Ono, et al. (2004) Angew Chem 43:4300-4302; Lee, et al. (1993) Biochem Cell Biol 71:162-168; Loakes, et al. (2009), Chem Commun 4619-4631; and Seo, et al. (2009) J Am Chem Soc 131:3246-3252, all incorporated herein by reference in their entireties for all purposes). Other types of modifications include, e.g, a nick, a missing base (e.g., apurinic or apyridinic sites), a pyrimidine dimer (e.g., thymine dimer or cyclobutane pyrimidine dimer), a cis-platin crosslinking, oxidation damage, hydrolysis damage, other methylated bases, bulky DNA base adducts, photochemistry reaction products, interstrand crosslinking products, mismatched bases, and other types of "damage" to the nucleic acid. As such, certain embodiments described herein refer to "damage" and such damage is also considered a modification of the nucleic acid in accordance with the present invention. Modified nucleotides can be caused by exposure of the DNA to radiation (e.g., UV), carcinogenic chemicals, crosslinking agents (e.g., formaldehyde), certain enzymes (e.g., nickases, glycosylases, exonucleases, methylases, other nucleases, etc.), viruses, toxins and other chemicals, thermal disruptions, and the like. *In vivo*, DNA damage is a major source of mutations leading to various diseases including cancer, cardiovascular disease, and nervous system diseases (see, e.g., Lindahl, T. (1993) Nature 362(6422): 709-15, which is incorporated herein by reference in its entirety for all purposes). The methods and systems provided herein can also be used to detect various conformations of DNA, in particular, secondary structure forms such as hairpin loops, stem-loops, internal loops, bulges, pseudoknots, base-triples, and the like; and are also useful for detection of agents interacting with the nucleic acid, e.g., bound proteins or other moieties.

[0052] In certain aspects, methods, compositions, and systems for detection and/or reversal of modifications in a template for single-molecule sequencing are provided, as well as determination of their location (i.e. "mapping") within a nucleic acid molecule. In certain

preferred embodiments, high-throughput, real-time, single-molecule, template-directed sequencing assays are used to detect the presence of such damage sites and to determine their location on the DNA template, e.g., by monitoring the progress and/or kinetics of a polymerase enzyme processing the template. For example, when a polymerase enzyme encounters certain types of damage or other modifications in a DNA template, the progress of the polymerase can be temporarily or permanently blocked, e.g., resulting in a paused or dissociated polymerase. As such, the detection of a pause in or termination of nascent strand synthesis is indicative of the presence of such damage or lesion. By analysis of the sequence reads produced prior to the pause or stop in synthesis, and alternatively or additionally after reinitiation of synthesis, one can map the site of the damage or lesion on the template. Since different types of lesions can have different effects on the progress of the polymerase on the substrate, in certain cases the behavior of the polymerase on the template not only informs as to where the lesion occurs, but also what type of lesion is present. Further, in certain embodiments a modification may be bypassed by incorporation of a non-nucleotide binding partner with the lesion in the template strand. For example, abasic sites (e.g., produced by glycosylases) can be "paired" with pyrenes or other similar analogs. (See, e.g., Matray, et al. (1999) Nature 399(6737): 704-8). Such an analog can also be labeled with a detectable label optically distinguishable from those on the nucleotides in the reaction mixture to allow optical detection of its incorporation. Certain aspects of the invention provide a means for reversing such modifications in real time, thereby allowing reinitation of the sequencing reaction and continued generation of sequence information for the template nucleic acid. Such methods can additionally be used to study effects of various agents (e.g., drugs, chemicals, enzymes, etc.) and reaction conditions on the creation and/or repair of such lesions and/or damage, as described elsewhere herein. These and other aspects of the invention are described in greater detail in the description and examples that follow.

II. Single Molecule Sequencing

[0053] In certain aspects of the invention, single molecule real time sequencing systems are applied to the detection of modified nucleic acid templates through analysis of the sequence and/or kinetic data derived from such systems. In particular, modifications in a template nucleic acid strand alter the enzymatic activity of a nucleic acid polymerase in various ways, e.g., by increasing the time for a bound nucleotide to be incorporated and/or increasing the time between incorporation events. In certain embodiments, polymerase activity is detected using a single molecule nucleic acid sequencing technology. In certain embodiments, polymerase activity is detected using a nucleic acid sequencing technology that detects incorporation of nucleotides into a nascent strand in real time. In preferred embodiments, a single molecule nucleic acid

sequencing technology is capable of real-time detection of nucleotide incorporation events. Such sequencing technologies are known in the art and include, e.g., the SMRT™ sequencing and nanopore sequencing technologies. For more information on nanopore sequencing, see, e.g., U.S. Patent No. 5,795,782; Kasianowicz, et al. (1996) Proc Natl Acad Sci USA 93(24):13770-3; Ashkenas, et al. (2005) Angew Chem Int Ed Engl 44(9):1401-4; Howorka, et al. (2001) Nat Biotechnology 19(7):636-9; and Astier, et al. (2006) J Am Chem Soc 128(5):1705-10, all of which are incorporated herein by reference in their entireties for all purposes. With regards to nucleic acid sequencing, the term "template" refers to a nucleic acid molecule subjected to template-directed synthesis of a nascent strand. A template may comprise DNA, RNA, or mimetics or derivatives thereof. Further, a template may be single-stranded, double-stranded, or may comprise both single- and double-stranded regions. A modification in a double-stranded template may be in the strand complementary to the newly synthesized nascent strand, or may by in the strand identical to the newly synthesized strand, i.e., the strand that is displaced by the polymerase.

[0054] The preferred direct methylation sequencing described herein may generally be carried out using single molecule real time sequencing systems, i.e., that illuminate and observe individual reaction complexes continuously over time, such as those developed for SMRT™ DNA sequencing (see, e.g., P. M. Lundquist, et al., *Optics Letters* **2008**, *33*, 1026, which is incorporated herein by reference in its entirety for all purposes). The foregoing SMRT™ sequencing instrument generally detects fluorescence signals from an array of thousands of ZMWs simultaneously, resulting in highly parallel operation. Each ZMW, separated from others by distances of a few micrometers, represents an isolated sequencing chamber.

[0055] Detection of single molecules or molecular complexes in real time, e.g., during the course of an analytical reaction, generally involves direct or indirect disposal of the analytical reaction such that each molecule or molecular complex to be detected is individually resolvable. In this way, each analytical reaction can be monitored individually, even where multiple such reactions are immobilized on a single substrate. Individually resolvable configurations of analytical reactions can be accomplished through a number of mechanisms, and typically involve immobilization of at least one component of a reaction at a reaction site. Various methods of providing such individually resolvable configurations are known in the art, e.g., see European Patent No.1105529 to Balasubramanian, et al.; and Published International Patent Application No. WO 2007/041394, the full disclosures of which are incorporated herein by reference in their entireties for all purposes. A reaction site on a substrate is generally a location on the substrate at which a single analytical reaction is performed and monitored, preferably in real time. A reaction site may be on a planar surface of the substrate, or may be in an aperture in

the surface of the substrate, e.g., a well, nanohole, or other aperture. In preferred embodiments, such apertures are "nanoholes," which are nanometer-scale holes or wells that provide structural confinement of analytic materials of interest within a nanometer-scale diameter, e.g., ~1-300 nm. In some embodiments, such apertures comprise optical confinement characteristics, such as zero-mode waveguides, which are also nanometer-scale apertures and are further described elsewhere herein. Typically, the observation volume (i.e., the volume within which detection of the reaction takes place) of such an aperture is at the attoliter (10^{-18} L) to zeptoliter (10^{-21} L) scale, a volume suitable for detection and analysis of single molecules and single molecular complexes.

[0056] The immobilization of a component of an analytical reaction can be engineered in various ways. For example, an enzyme (e.g., polymerase, reverse transcriptase, kinase, etc.) may be attached to the substrate at a reaction site, e.g., within an optical confinement or other nanometer-scale aperture. In other embodiments, a substrate in an analytical reaction (for example, a nucleic acid template, e.g., DNA, RNA, or hybrids, analogs, and mimetics thereof, or a target molecule for a kinase) may be attached to the substrate at a reaction site. Certain embodiments of template immobilization are provided, e.g., in U.S. Patent Application No. 12/562,690, filed September 18, 2009 and incorporated herein by reference in its entirety for all purposes. One skilled in the art will appreciate that there are many ways of immobilizing nucleic acids and proteins into an optical confinement, whether covalently or non-covalently, via a linker moiety, or tethering them to an immobilized moiety. These methods are well known in the field of solid phase synthesis and micro-arrays (Beier et al., Nucleic Acids Res. 27:1970-1-977 (1999)). Non-limiting exemplary binding moieties for attaching either nucleic acids or polymerases to a solid support include streptavidin or avidin/biotin linkages, carbamate linkages, ester linkages, amide, thiolester, (N)-functionalized thiourea, functionalized maleimide, amino, disulfide, amide, hydrazone linkages, among others. Antibodies that specifically bind to one or more reaction components can also be employed as the binding moieties. In addition, a silyl moiety can be attached to a nucleic acid directly to a substrate such as glass using methods known in the art.

[0057] In some embodiments, a nucleic acid template is immobilized onto a reaction site (e.g., within an optical confinement) by attaching a primer comprising a complementary region at the reaction site that is capable of hybridizing with the template, thereby immobilizing it in a position suitable for monitoring. In certain embodiments, an enzyme complex is assembled in an optical confinement, e.g., by first immobilizing an enzyme component. In other embodiments, an enzyme complex is assembled in solution prior to immobilization. Where desired, an enzyme or other protein reaction component to be immobilized may be modified to contain one or more

epitopes for which specific antibodies are commercially available. In addition, proteins can be modified to contain heterologous domains such as glutathione S-transferase (GST), maltose-binding protein (MBP), specific binding peptide regions (see e.g., U.S. Patent Nos. 5,723,584, 5,874,239 and 5,932,433), or the Fc portion of an immunoglobulin. The respective binding agents for these domains, namely glutathione, maltose, and antibodies directed to the Fc portion of an immunoglobulin, are available and can be used to coat the surface of an optical confinement of the present invention.The binding moieties or agents of the reaction components they immobilize can be applied to a support by conventional chemical techniques which are well known in the art. In general, these procedures can involve standard chemical surface modifications of a support, incubation of the support at different temperature levels in different media comprising the binding moieties or agents, and possible subsequent steps of washing and cleaning.

[0058] In some embodiments, a substrate comprising an array of reaction sites is used to monitor multiple biological reactions, each taking place at a single one of the reaction sites. Various means of loading multiple biological reactions onto an arrayed substrate are known to those of ordinary skill in the art and are described further, e.g., in USSN 61/072,641, incorporated herein by reference in its entirety for all purposes. For example, basic approaches include: creating a single binding site for a reaction component at the reaction site; removing excess binding sites at the reaction site via catalytic or secondary binding methods; adjusting the size or charge of the reaction component to be immobilized; packaging or binding the reaction component within (or on) a particle (e.g., within a viral capsid), where a single such particle fits into the relevant reaction site (due to size or charge of the particle and/or observation volume); using non-diffusion limited loading; controllably loading the reaction component (e.g., using microfluidic or optical or electrical control); sizing or selecting charges in the reaction sites/observation volumes (e.g., the sizes of optical confinements in an array) to control which reaction components will fit (spatially or electrostatically) into which reaction sites/observation volumes; iterative loading of reaction components, e.g., by masking active sites between loading cycles; enriching the activity of the reaction components that are loaded; using self-assembling nucleic acids to sterically control loading; adjusting the size of the reaction site/observation volume; and many others. Such methods and compositions provide for the possibility of completely loading single-molecule array reaction sites (instead of about 30% of such sites as occurs in "Poisson limited" loading methods) with single reaction components (e.g., molecular complexes).

[0059] In preferred aspects, the methods, compositions, and systems provided herein utilize optical confinements to facilitate single molecule resolution of analytical reactions. In preferred

embodiments, such optical confinements are configured to provide tight optical confinement so only a small volume of the reaction mixture is observable. Some such optical confinements and methods of manufacture and use thereof are described at length in, e.g., U.S. Patent. Nos. 7,302,146. 7,476,503, 7,313,308, 7,315,019, 7,170,050, 6,917,726, 7,013,054, 7,181,122, and 7,292,742; U.S. Patent Publication Nos. 20080128627, 20080152281, and 200801552280; and U.S.S.N. 11/981,740 and 12/560,308, all of which are incorporated herein by reference in their entireties for all purposes.

[0060] Where reaction sites are located in optical confinements, the optical confinements can be further tailored in various ways for optimal confinement of an analytical reaction of interest. In particular, the size, shape, and composition of the optical confinement can be specifically designed for containment of a given enzyme complex and for the particular label and illumination scheme used.

[0061] In certain preferred embodiments of the invention, single-molecule real-time sequencing systems already developed are applied to the detection of modified nucleic acid templates through analysis of the sequence and kinetic data derived from such systems. As described below, methylated cytosine and other modifications in a template nucleic acid will alter the enzymatic activity of a polymerase processing the template nucleic acid. In certain embodiments, polymerase kinetics in addition to sequence read data are detected using a single molecule nucleic acid sequencing technology, e.g., the SMRT™ sequencing technology developed by Pacific Biosciences (Eid, J. et al. (2009) *Science* **2009**, *323*, 133, the disclosure of which is incorporated herein by reference in its entirety for all purposes). This technique is capable of long sequencing reads and provides high-throughput methylation profiling even in highly repetitive genomic regions, facilitating *de novo* sequencing of modifications such as methylated bases. SMRT™ sequencing systems typically utilize state-of-the-art single-molecule detection instruments, production-line nanofabrication chip manufacturing, organic chemistry, protein mutagenesis, selection and production facilities, and software and data analysis infrastructures.

[0062] Certain preferred methods of the invention employ real-time sequencing of single DNA molecules (Eid, et al., *supra*), with intrinsic sequencing rates of several bases per second and average read lengths in the kilobase range. In such sequencing, sequential base additions catalyzed by DNA polymerase into the growing complementary nucleic acid strand are detected with fluorescently labeled nucleotides. The kinetics of base additions and polymerase translocation are sensitive to the structure of the DNA double-helix, which is impacted by the presence of base modifications, e.g, 5-MeC and other perturbations (secondary structure, bound agents, etc.) in the template. By monitoring the activity of DNA polymerase during sequencing,

sequence read information and base modifications can be simultaneously detected. Long, continuous sequence reads that are readily achievable using SMRT™ sequencing facilitate modification (e.g., methylation) profiling in low complexity regions that are inaccessible to some technologies, such as certain short-read sequencing technologies. Carried out in a highly parallel manner, methylomes can be sequenced directly, with single base-pair resolution and high throughput.

[0063] The principle of SMRT™ sequencing is illustrated in Figure 1. Two important technology components of certain embodiments of this process are: (i) optical confinement technology that allows single-molecule detection at concentrations of labeled nucleotides relevant to the enzyme, and (ii) phospholinked nucleotides that enable observation of uninterrupted polymerization.

[0064] In preferred embodiments, optical confinements are ZMW nanostructures, preferably in an arrayed format. Typically, ZMWs arrays comprise dense arrays of holes, ~100 nm in diameter, fabricated in a ~100 nm thick metal film deposited on a transparent substrate (e.g., silicon dioxide). These structures are further described in the art, e.g., in M. J. Levene, et al., *Science* **2003**, *299*, 682 ; and M. Foquet, et al., *J. Appl. Phys.* **2008**, *103*, 034301, the disclosures of which are incorporated herein by reference in their entireties for all purposes. Each ZMW becomes a nanophotonic visualization chamber for recording an individual polymerization reaction, providing a detection volume of just 100 zeptoliters (10^{-21} liters). This volume represents a ~1000-fold improvement over diffraction-limited confocal microscopy, facilitating observation of single incorporation events against the background created by the relatively high concentration of fluorescently labeled nucleotides. Polyphosphonate and silane-based surface coatings mediate enzyme immobilization to the transparent floor of the ZMW while blocking non-specific attachments to the metal top and side wall surfaces (Eid, et al., *supra*; and J. Korlach, et al., *Proc Natl Acad Sci U S A* **2008**, *105*, 1176, the disclosures of which are incorporated herein by reference in their entireties for all purposes). While certain methods described herein involve the use of ZMW confinements, it will be readily understood by those of ordinary skill in the art upon review of the teachings herein that these methods may also be practiced using other reaction formats, e.g., on planar substrates or in nanometer-scale apertures other than zero-mode waveguides. (See, e.g., U.S.S.N. 12/560,308, filed September 15, 2009; and U.S. Patent Publication No. 20080128627, incorporated herein *supra*.)

[0065] The second important component is phospholinked nucleotides for which a detectable label (e.g., comprising a fluorescent dye) is attached to the terminal phosphate rather than the base (Figure 1). (See, e.g., J. Korlach, et al., *Nucleos. Nucleot. Nucleic Acids* **2008**, *27*, 1072, which is incorporated herein by reference in its entirety for all purposes.) 100% replacement of

unmodified nucleotides by phospholinked nucleotides is performed, and the enzyme cleaves away the label as part of the incorporation process, leaving behind a completely natural, double-stranded nucleic acid product. Each of the four different nucleobases is labeled with a distinct detectable label to discriminate base identities during incorporation events, thus enabling sequence determination of the complementary DNA template. During incorporation, the enzyme holds the labeled nucleotide in the ZMW's detection volume for tens of milliseconds, orders of magnitude longer than the average diffusing nucleotide is present. Signal (e.g., fluorescence) is emitted continuously from the detectable label during the duration of incorporation, causing a detectable pulse of increased fluorescence in the corresponding color channel. The pulse is terminated naturally by the polymerase releasing the pyrophosphate-linker-label group. The polymerase then translocates to the next base, and the process repeats. As shown in Figure 1A, single DNA polymerase molecules with bound DNA template are attached to a substrate, e.g., at the bottom of each zero-mode waveguide. Polymerization of the complementary DNA strand is observed in real time by detecting fluorescently labeled nucleotides. Reactions steps involved in SMRT™ sequencing are as follows: Step 1: The DNA template/primer/polymerase complex is surrounded by diffusing fluorescently labeled nucleotides which probe the active site. Step 2: A labeled nucleotide makes a cognate binding interaction with the next base in the DNA template that lasts for tens of milliseconds, during which fluorescence is emitted continuously. Step 3: The polymerase incorporates the nucleotide into the growing nucleic acid chain, thereby cleaving the α-β phosphodiester bond, followed by release of the nucleotide. Steps 4-5: The process repeats. A prophetic trace is shown in Figure 1B that comprises each step shown in 1A. At steps 2 and 4, a fluorescent signal is emitted during binding and incorporation of a nucleotide into the growing nucleic acid chain, and monitoring of these fluorescent signals provides a sequence of nucleotide incorporations that can be used to derive the sequence of the template nucleic acid. For example, a 5'-G-A-3' sequence in the growing chain indicates a 5'-T-C-3' sequence in the complementary template strand.

[0066] As described above, reaction data is indicative of the progress of a reaction and can serve as a signal for the presence of a modification in the template nucleic acid. Reaction data in single molecule sequencing reaction reactions using fluorescently labeled bases is generally centered around characterization of detected fluorescence pulses, a series of successive pulses ("pulse trace" or one or more portions thereof), and other downstream statistical analyses of the pulse and trace data. Fluorescence pulses are characterized not only by their spectrum, but also by other metrics including their duration, shape, intensity, and by the interval between successive pulses (see, e.g., Eid, et al., *supra*; and U.S. Patent Publication No. 20090024331, incorporated herein by reference in its entirety for all purposes). While not all of these metrics are generally

required for sequence determination, they add valuable information about the processing of a template, e.g., the kinetics of nucleotide incorporation and DNA polymerase processivity and other aspects of the reaction. Further, the context in which a pulse is detected (i.e., the one or more pulses that precede and/or follow the pulse) can contribute to the identification of the pulse. For example, the presence of certain modifications alters not only the processing of the template at the site of the modification, but also the processing of the template upstream and/or downstream of the modification. For example, the presence of modified bases in a template nucleic acid has been shown to change the width of a pulse and/or the interpulse duration (IPD), either at the modified base or at one or more positions proximal to it. A change in pulse width may or may not be accompanied by a change in IPD. Figure 2 provides illustrative examples of various types of reaction data in the context of a pulse trace including IPD, pulse width (PW), pulse height (PH), and context. Figure 2A illustrates these reaction data on a pulse trace generated on an unmodified template, and Figure 2B illustrates how the presence of a modification (5-MeC) can elicit a change in one of these reaction data (IPD) to generate a signal (increased IPD) indicative of the presence of the modification.

[0067] In yet further embodiments, reaction data is generated by analysis of the pulse and trace data to determine error metrics for the reaction. Such error metrics include not only raw error rate, but also more specific error metrics, e.g., identification of pulses that did not correspond to an incorporation event, incorporations that were not accompanied by a detected pulse, incorrect incorporation events, and the like. Any of these error metrics, or combinations thereof, can serve as a signal indicative of the presence of one or more modifications in the template nucleic acid. In some embodiments, such analysis involves comparison to a reference sequence and/or comparison to replicate sequence information from the same or an identical template, e.g., using a standard or modified multiple sequence alignment. Certain types of modifications cause an increase in one or more error metrics. For example, some modifications can be "paired" with more than one type of incoming nucleotide or analog thereof, so replicate sequence reads for the region comprising the modification will show variable base incorporation opposite such a modification. Such variable incorporation is thereby indicative of the presence of the modification. Certain types of modifications cause an increase in one or more error metrics proximal to the modification, e.g., immediately upstream or downstream. The error metrics at a locus or within a region of a template are generally indicative of the type of modification(s) present at that locus or in that region of the template, and therefore serve as a signal of such modification(s). In preferred embodiments, at least some reaction data is collected in real time during the course of the reaction, e.g., pulse and/or trace characteristics.

[0068] Although described herein primarily with regards to fluorescently labeled nucleotides, other types of detectable labels and labeling systems can also be used with the methods, compositions, and systems described herein including, e.g., quantum dots, surface enhanced Raman scattering particles, scattering metallic nanoparticles, FRET systems, intrinsic fluorescence, non-fluorescent chromophores, and the like. Such labels are generally known in the art and are further described in Provisional U.S. Patent Application No. 61/186,661, filed June 12, 2009; U.S. Patent Nos. 6,399,335, 5,866,366, 7,476,503, and 4,981,977; U.S. Patent Pub. No. 2003/0124576; U.S.S.N. 61/164567; WO 01/16375; Mujumdar, et al Bioconjugate Chem. 4(2):105-111, 1993; Ernst, et al, Cytometry 10:3-10, 1989; Mujumdar, et al, Cytometry 10:1119, 1989; Southwick, et al, Cytometry 11:418-430, 1990; Hung, et al, Anal. Biochem. 243(1):15-27, 1996; Nucleic Acids Res. 20(11):2803-2812, 1992; and Mujumdar, et al, Bioconjugate Chem. 7:356-362, 1996; Intrinsic Fluorescence of Proteins, vol. 6, publisher: Springer US, ©2001; Kronman, M.J. and Holmes, L.G. (2008) Photochem and Photobio 14(2): 113-134; Yanushevich, Y.G., et al. (2003) Russian J. Bioorganic Chem 29(4) 325-329; and Ray, K., et al. (2008) J. Phys. Chem. C 112(46): 17957-17963, all of which are incorporated herein by reference in their entireties for all purposes. Many such labeling groups are commercially available, e.g., from the Amersham Biosciences division of GE Healthcare, and Molecular Probes/Invitrogen Inc. (Carlsbad, CA)., and are described in 'The Handbook — A Guide to Fluorescent Probes and Labeling Technologies, Tenth Edition' (2005) (available from Invitrogen, Inc./Molecular Probes and incorporated herein in its entirety for all purposes). Further, a combination of the labeling strategies described herein and known in the art for labeling reaction components can be used.

[0069] Various strategies, methods, compositions, and systems are provided herein for detecting modifications in a nucleic acid, e.g., during real-time nascent strand synthesis. For example, since DNA polymerases can typically bypass 5-MeC in a template nucleic acid and properly incorporate a guanine in the complementary strand opposite the 5-MeC, additional strategies are desired to detect such altered nucleotides in the template. Various such strategies are provided herein, such as, e.g., a) modification of the polymerase to introduce an specific interaction with the modified nucleotide; b) detecting variations in enzyme kinetics, e.g., pausing; c) use of a detectable and optionally modified nucleotide analog that specifically base-pairs with the modification and is potentially incorporated into the nascent strand; d) chemical treatment of the template prior to sequencing that specifically alters 5-MeC sites in the template; e) use of a protein that specifically binds to the modification in the template nucleic acid, e.g., delaying or blocking progression of a polymerase during replication; and f) use of sequence context (e.g., the higher frequency of 5-MeC nucleotides in CpG islands) to focus modification

detection efforts on regions of the template that are more likely to contain such a modification (e.g., GC-rich regions for 5-MeC detection). These strategies may be used alone or in combination to detect 5-MeC sites in a template nucleic acid during nascent strand synthesis.

III. Polymerase Modifications

[0070] Various different polymerases may be used in template-directed sequence reactions, e.g., those described at length, e.g., in U.S. Pat. No. 7,476,503, the disclosure of which is incorporated herein by reference in its entirety for all purposes. In brief, the polymerase enzymes suitable for the present invention can be any nucleic acid polymerases that are capable of catalyzing template-directed polymerization with reasonable synthesis fidelity. The polymerases can be DNA polymerases or RNA polymerases (including, e.g., reverse transcriptases), a thermostable polymerase or a thermally degradable polymerase, wildtype or modified. In some embodiments, the polymerases exhibit enhanced efficiency as compared to the wildtype enzymes for incorporating unconventional or modified nucleotides, *e.g.*, nucleotides linked with fluorophores. In certain preferred embodiments, the methods are carried out with polymerases exhibiting a high degree of processivity, *i.e.*, the ability to synthesize long stretches (e.g., over about 10 kilobases) of nucleic acid by maintaining a stable nucleic acid/enzyme complex. In certain preferred embodiments, sequencing is performed with polymerases capable of rolling circle replication. A preferred rolling circle polymerase exhibits strand-displacement activity, and as such, a single circular template can be sequenced repeatedly to produce a sequence read comprising multiple copies of the complement of the template strand by displacing the nascent strand ahead of the translocating polymerase. Since the methods of the invention can increase processivity of the polymerase by removing lesions that block continued polymerization, they are particularly useful for applications in which a long nascent strand is desired, e.g. as in the case of rolling-circle replication. Non-limiting examples of rolling circle polymerases suitable for the present invention include but are not limited to T5 DNA polymerase, T4 DNA polymerase holoenzyme, phage M2 DNA polymerase, phage PRD1 DNA polymerase, Klenow fragment of DNA polymerase, and certain polymerases that are modified or unmodified and chosen or derived from the phages Φ29 (Phi29), PRD1, Cp-1, Cp-5, Cp-7, Φ15, Φ1, Φ21, Φ25, BS 32 L17, PZE, PZA, Nf, M2Y (or M2), PR4, PR5, PR722, B103, SF5, GA-1, and related members of the *Podoviridae* family. In certain preferred embodiments, the polymerase is a modified Phi29 DNA polymerase, e.g., as described in U.S. Patent Publication No. 20080108082, incorporated herein by reference in its entirety for all purposes. Additional polymerases are provided, e.g., in U.S.S.N. 11/645,125, filed December 21, 2006; 11/645,135, filed December 21, 2006; 12/384,112, filed March 30, 2009; and 61/094,843, filed September 5,

2008; as well as in U.S. Patent Publication No. 20070196846, the disclosures of which are incorporated herein by reference in their entireties for all purposes.

[0071] Further optimization is achieved through improvement of enzyme kinetics, either through the screening of polymerase libraries and/or the engineering of polymerases, which include, e.g., DNA polymerases, RNA polymerases, reverse transcriptases, and the like. In particular, DNA polymerases may be screened to identify those that have desirable properties for detection of nucleic acid modifications described herein. Further, polymerases can be engineered through directed mutagenesis of one or more residues involved in various aspects of template-directed nascent strand synthesis. For example, careful examination of the crystal structure of the polymerase-DNA-nucleotide complex for certain polymerase enzymes has shown that the polymerase rotates and flips out the base on the single-stranded region of the template DNA that is adjacent to the active site in the 5' direction (i.e., the base in the "-1" position). During the subsequent DNA translocation process, this base is flipped into the active site. As such, amino acids that interact with a modified base in the -1 position or during the subsequent tranlocation can be altered or substituted to increase the enzyme's sensitivity to the modified base. In fact, any protein residues that come into close contact with a modification in the template are candidates for substitution or alteration. For example, family B polymerases mostly contain replicative polymerases and include the major eukaryotic DNA polymerases α, δ, ε, and also DNA polymerase ζ. Family B also includes DNA polymerases encoded by some bacteria and bacteriophages, e.g., T4, Phi29, and RB69 bacteriophages. Most family B polymerases share common structural features for DNA binding, and the residues along the DNA primer-template junction and the residues around the base binding pocket at the -1 location (pre-insertion position) can be mutated and the resulting mutants screened for enhanced response to a modification of interest. Specifically, when 5-MeC is in the "flipped-out" (-1) position, it is surrounded by several Φ29 polymerase amino acid residues, such as K392 and K422 which are positioned close to the methyl group (Figure 3). Mutations such as K392R/W/M and K422R/W/M that substitute the native lysine residue with amino acids with larger side chains (e.g., arginine, tryptophan, or methionine) may increase the polymerase's sensitivity to modified bases, potentially delaying the translocation step and slowing incorporation of the complementary dGTP. This is schematically illustrated with reference to Figure 3, which shows a structural model of 5-MeC positioned one base in the 5' direction relative to the Φ29 DNA polymerase active site. As shown, two polymerase residues, K392 and K422, are close to the methyl group of 5-MeC. These two residues are potential targets for site-specific mutagenesis to amino acids with larger side chains that will interact sterically with the methyl group. 5-MeC is

shown as the chemical structure in the right center, while the K392 and K422 residues are left center.

[0072] Any residues that come in close contact with a modified base are candidates for mutation. For example, using molecular morphing and energy minimization to model the translocation path of 5-MeC, a number of protein residues in the Φ29 polymerase have been identified within 5 Å of the methyl moiety as it is being flipped into the active site during the translocation step. The following groups of residues (listed in the order of the translocation path) are targets of mutations to residues with larger side chains: I93, M188, K392, V399, T421, K422; S95, Y101, M102; Q99, L123, K124, T189, A190; G191, S388; P127, L384, N387, S388; and L389, Y390, G391. In particular, I93Y and V399Y may introduce a 5-methylcytosine specific binding region, analogous to those shown by the crystal structures of the SRA/5-methylcytosine binding complex. For example, see G. V. Avvakumov, et al., *Nature* **2008**, *455*, 822; and H. Hashimoto, et al., *Nature* **2008**, *455*, 826, the disclosures of both of which are incorporated herein by reference in their entireties for all purposes. Although the residues identified above are specific to the Φ29 polymerase, one of ordinary skill will readily recognize that the structural similarity between the family B polymerases, and to a lesser extent family A polymerases and other polymerases, allows identification of homologous positions on related polymerases as targets for mutation based on the teachings herein.

[0073] In addition to the foregoing, additional improvements are derived from a molecular evolution program using these polymerases to enhance their ability to sense 5-MeC. Such programs have already been used to successfully improve large numbers of different enzymes for a variety of applications, including improving DNA polymerases for sequencing. Such methods may include diversification of the amino acid sequence space by mutagenic PCR and DNA shuffling, and/or yeast displays for expression and selection (see, e.g., S. A. Gai, et al., *Curr Opin Struct Biol* **2007**, *17*, 467; and D. Lipovsek, et al., *Chem Biol* **2007**, *14*, 1176, which are incorporated herein by reference in their entireties for all purposes, in which ~10^4 copies of a recombinant protein are displayed on the surface of a single yeast cell carrying the transgene for the protein. The genotype-phenotype linkage is provided by the yeast cell, but no protein purification is necessary as the displayed proteins have the same properties as bulk solutions of polymerase. With the available infrastructure in house, this program can be initiated without startup costs as soon as polymerase candidates emerge.

IV. Secondary Structure Detection

[0074] During single molecule sequencing as described *supra*, an otherwise highly processive trace is sometimes interrupted by a long pause. Such pausing can be caused by secondary structure, e.g., a hairpin loop, in the template strand. In certain aspects, the invention provides methods for not only identifying secondary structure in a template nucleic acid, but also for improving the overall accuracy of single molecule sequencing.

[0075] In certain embodiments, a sequencing read or "trace" is generated by subjecting a template nucleic acid to a real-time, template-directed sequencing reaction. The trace is examined to identify long pauses by finding portions of the trace at which the interpulse duration (IPD) is significantly longer than the average IPD. For example, find pauses that are at least 2-, 3-, 5-, 10-, or 20-times longer than the average IPD. In some embodiments, an IPD averaged over a few neighboring bases is used, in other embodiments an IPD averaged over a window of about 20, 30, 50, 70, or 100 bases is used, and in yet other embodiments, an IPD averaged over all or substantially all of the template is used.

[0076] The sequence reads generated before and after the pause are analyzed within about a 20-, 30-, 50-, 70-, or 100-base window centered on the pause, and regions that flank the pause site and are complementary to one another are identified. Based upon the complementary sequences, their spacing, and other known factors that impact secondary structure formation (e.g., GC content, pH, salt concentration, and the like), the probability of a hairpin loop at that location in the template is determined. If this probability is high, the sequence reads flanking the pause site are re-examined to identify basecalls that do not match the complementarity of the hairpin, e.g., a non-complementary basecall or missing basecall within a stretch of complementary basecalls. Such non-complementary or missing basecalls have a higher probability of being errors than basecalls in the region that do not interrupt the complementarity between the regions upstream and downstream of the pause site. As such, the basecalls at these positions are reevaluated to determine if the initial basecall was erroneous. Further, knowledge of a given template's propensity for forming secondary structures that interfere with processivity of a polymerase can be used in future rounds of template-directed sequencing of the template to better call base positions in the vicinity of the interfering secondary structure, thereby improving accuracy of basecalls in the future rounds.

[0077] Additionally, since the duration of the pause is likely related to the strength of the secondary structure formed within the template, the duration can be used as a metric in determining the type, size, compositions, and likelihood of a secondary structure in the template molecule. In addition, for applications in which a single template is repeatedly subjected to template-directed synthesis, the replicate sequence reads that are generated are compared to one

another to determine if a given portion of the template consistently produces a pause in the synthesis reaction, which provides further evidence that the pause is due to the sequence context, e.g., secondary structure spontaneously forming in the template.

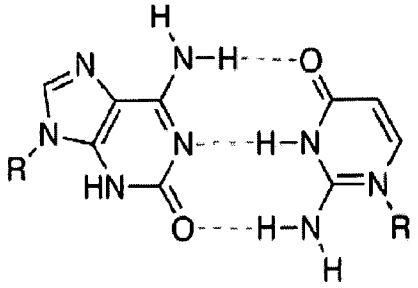
IV. Modified and Non-natural Nucleotide Analogs and Base Pairing

[0078] In certain aspects, methods, compositions, and systems are provided that utilize modified and/or non-natural nucleotide analogs and/or base pairing. For example, certain non-natural nucleotide analogs can be incorporated by a polymerase into a nascent strand opposite a modification, e.g., missing or damaged base. In certain embodiments, such non-natural nucleotide analogs are detectably labeled such that their incorporation can be distinguished from incorporation of a natural nucleotide or nucleotide analog, e.g., during template-directed nascent strand synthesis. This strategy allows real-time sequencing that generates reads that not only provide base sequence information for native bases in the template, but also modified bases without requiring further modifications to the standard methods (Eid, et al, *supra*). This method facilitates modification profiling in the absence of repeated sequencing of each DNA template, and is particularly well suited to *de novo* applications. In certain embodiments, the modified or non-natural nucleotide analogs are not incorporatable into the nascent strand and the polymerase can bypass the modification using a native nucleotide or nucleotide analog, which may or may not be labeled. Since the modified or non-natural analog has a higher affinity for the modification than a native analog, it will bind to the polymerase complex multiple times before a native analog is incorporated, resulting in multiple signals for a single incorporation event, and thereby increasing the likelihood of accurate detection of the modification. Similar methods for use in sequencing unmodified template nucleic acids are described in greater detail in U.S.S.N. 61/186,661, filed June 12, 2009 and incorporated herein by reference in its entirety for all purposes.

[0079] Since 5-MeC retains Watson-Crick hydrogen bonding with guanine, an incoming guanine nucleotide analog can be used to detect 5-MeC in the template strand. For example, a guanine nucleotide analog can be constructed to cross the major groove and sense the modified cytosine therein. In particular embodiments, a fused aromatic ring is linked to the N7 atom of the guanine of the nucleotide analog. This modified guanine nucleotide analog can "sense" the methyl group of 5-MeC and affect the base-pairing as compared to an unmodified guanine nucleotide analog. Such guanine nucleotide analogs are further described elsewhere, e.g., in International Application Pub. No. WO/2006/005064 and U.S. Patent No. 7,399,614. Similar modifications can be made to nucleotide analogs appropriate for SMRT™ sequencing applications, e.g., those with terminal-phosphate labels, e.g., as described in U.S. Patent Nos.

7,056,661 and 7,405,281; U.S. Patent Pub. Nos. 20070196846 and 20090246791; and U.S.S.N. 12/403,090, all of which are incorporated herein by reference in their entireties for all purposes. In certain embodiments, 5-MeC detection may be carried out using a modified guanine nucleotide analog described above that carries a detectable label that is distinguishable from detectable labels on other reaction components, e.g., other nucleotide analogs being incorporated. Such a strategy allows 5-MeC detection by observation of a signal, rather than or in addition to altered polymerase kinetics, which facilitates methylation profiling even in the absence of redundant or replicate sequencing of the template.

[0080] Certain embodiments use other non-natural base pairs that are orthogonal to the natural nucleobases pairs. For example, isoguanine (isoG) can be incorporated by a polymerase into DNA at sites complementary to isocytosine (isoC) or 5-methylisocytosine (^{Me}isoC), and vice versa, as shown by the following chemical structure and described in A.T. Krueger, et al., "Redesigning the Architecture of the Base Pair: Toward Biochemical and Biological Function of New Genetic Sets." *Chemistry & Biology* **2009**, *16*(3), 242, incorporated herein by reference in its entirety for all purposes.



iso-G *iso*-C

Other non-natural base pairs that are orthogonal to the natural nucleobases pairs can also be used, e.g., Im-N^O/Im-O^N or A*/T* (described further in J. D. Ahle, et al., *Nucleic Acids Res* **2005**, *33*(10), 3176; A.T. Krueger, et al., *supra*; and A.T. Krueger, et al., Curr Opinions in Chem Biology 2007, *11*(6), 588).

[0081] In certain embodiments, a nucleic acid modification to be detected by the methods herein is 7,8-dihydro-8-oxoguanine ("8-oxoG") (also known as 8-oxo-7,8-dihydroguanine, 8-oxoguanine, and 8-hydroxyguanine). 8-oxoG is the major oxidative DNA lesion found in human tissue. Due to the relatively subtle modification to guanine in 8-oxoG, it may be bypassed by replicative DNA polymerases, which preferentially incorporate an adenine nucleotide into the nascent nucleic acid strand at the position where the complementary cytosine should be incorporated, thereby resulting in a mutation in the nascent strand (see, e.g., Hsu, et al. (2004)

Nature 431(7005): 217-21; and Hanes, et al. (2006) J. Biol. Chem. 281:36241-8, which are incorporated herein by reference in their entireties for all purposes). As well as introducing mutations *in vivo*, the bypass of such lesions by a polymerase during template-dependent sequencing reactions introduces errors into the sequence reads generated, and the presence of the damaged guanine nucleotide can also cause base misalignment, potentially adding further errors into a resulting sequence read. DNA synthesis opposite an 8-oxoG lesion has relatively very low specificity (kcat/Km) that is about 10^6-fold lower than incorporating a C opposite an unmodified G. See, e.g., Hsu, et al., *supra*. Further, due to its very low redox potential 8-oxoG can be more easily oxidized than unmodified guanine, and the 8-oxoG oxidation products are very effective blockers of DNA polymerases. See, e.g., Duarte, et al. (1999) Nucleic Acids Res 27(2):496-502; and Kornyushyna, et al. (2002) Biochemistry 41(51): 15304-14, the disclosures of which are incorporated herein by reference in their entireties for all purposes.

[0082] It has been shown that 8-oxoG alters both k_{cat} and K_m of steady-state incorporation kinetics, which are likely to cause altered pulse widths and IPD before incorporation of a nucleotide (G or A) into the complementary position in the nascent strand during template-directed sequencing reactions (see, e.g., Hsu, et al. and Hanes, et al., *supra*). These altered kinetic characteristics can be used to detect 8-oxoG in a template nucleic acid during real-time sequencing reactions. Further, a circular template that comprises both complementary strands of a region of interest (e.g., as described in U.S.S.N. 12/383,855 and 12/413,258, both filed March 27, 2009 and incorporated herein by reference in their entireties for all purposes) can be used to repeatedly sequence both strands of a region of interest, thereby generating redundant sequence information that can be analyzed to statistically determine how often a given position in the template has an A-G mismatch as compared to how often the correct base is incorporated at that position. The redundant sequence information increases the accuracy of correctly calling a position as a G or an 8-oxoG. For example, if the mismatch rate is 100%, then if one detects an A at the position, but then a G at the complementary position, then it is highly likely that the A detected was Hoogsteen base pairing with an 8-oxoG in the template. This strategy is similar to detection of 5-MeC modifications that have been deaminated to uracil prior to sequencing, as described in greater detail below.

[0083] The mismatch incorporation rate opposite 8-oxoG sites, as well as the degree to which IPD and pulse width are affected by 8-oxoG depend on the type of polymerase used in the reaction (see, e.g., Hsu, et al. and Hanes, et al., *supra*). As such, polymerase mutants can be designed to have increased kinetic sensitivity to 8-oxoG, or increased/decreased misincorporation rate opposite an 8-oxoG. Methods for designing polymerases for various embodiments of the invention are known in the art and provided elsewhere herein. Further,

multiple binding events are very likely at the site of modification, resulting in one or more signals not associated with incorporation into the nascent strand, and these multiple binding events can also occur at positions proximal to the modification, e.g., continuing for a few bases after the site of damage. These additional signaling events would provide a robust indicator of the site of modification. In addition, multiple sequencing reads for the region of the template comprising the modification are expected to contain variable numbers of extra signaling events at or proximal to the modification. As such, comparison of this redundant sequence data will also facilitate identification of loci comprising the modification.

V. Chemical Modification of Template

[0084] Direct detection of modifications (e.g., methylated bases as described above) without pre-treatment of the DNA sample, has many benefits. Alternatively or additionally, complementary techniques may be employed, such as the use of non-natural or modified nucleotide analogs and/or base pairing described elsewhere herein. In general, such complementary techniques serve to enhance the detection of the modification, e.g., by amplifying a signal indicative of the modification. Further, while the methods described herein focus primarily on detection of 5-MeC nucleotides, it will be clear to those of ordinary skill in the art that these methods can also be extended to detection of other types of nucleotide modifications or damage. In addition, since certain sequencing technologies (e.g., SMRT™ sequencing) do not require amplification of the template, e.g., by PCR, other chemical modifications of the 5-MeC can be employed to facilitate detection of these modified nucleotides in the template. For example, the difference in redox potential between normal cytosine and 5-MeC can be used to selectively oxidize 5-MeC and further distinguish it from the nonmethylated base. Such methods are further described elsewhere, and include halogen modification (S. Bareyt, et al., *Angew Chem Int Ed Engl* **2008**, *47*(1), 181) and selective osmium oxidation (A. Okamoto, *Nucleosides Nucleotides Nucleic Acids* **2007**, *26*(10-12), 1601; and K. Tanaka, et al., *J Am Chem Soc* **2007**, *129*(17), 5612), and these references are incorporated herein by reference in their entireties for all purposes.

Glycosylase Modification

[0085] By way of example, DNA glycosylases are a family of repair enzymes that excise altered, damaged, or mismatched nucleotide residues in DNA while leaving the sugar-phosphate backbone intact. Additional information on glycosylase mechanisms and structures is provided in the art, e.g., in A. K. McCullough, et al., *Annual Rev of Biochem* **1999**, *68*, 255. In particular, four DNA glycosylases (ROS1, DME, DML2, and DML3) have been indentified in *Arabidopsis thaliana* that remove methylated cytosine from double-stranded DNA, leaving an abasic site.

(See, e.g., S. K. Ooi, et al., *Cell* **2008**, *133*, 1145, incorporated herein by reference in its entirety for all purposes.) Furthermore, it has been shown that a 5'-triphosphate derivative of the pyrene nucleoside (dPTP) is efficiently and specifically inserted by certain DNA polymerases into abasic DNA sites through steric complementarity. (See, e.g., T. J. Matray, et al., *Nature* **1999**, *399*(6737), 704, incorporated herein by reference in its entirety for all purposes.)

[0086] In certain embodiments of single-molecule, five-color DNA methylation sequencing, DNA glycosylase activity can be combined with polymerase incorporation of a non-natural nucleotide analog (e.g., a pyrene analog (dPTP) as shown in Figure 4). For example, in certain embodiments, methylated cytosines are excised from a DNA sample treated with an *Arabidopsis* DNA glycosylase. Covalent linkage of a fifth fluorophore to the terminal phosphate of dPTP allows detection of abasic sites during polymerase-mediated DNA synthesis.

[0087] Figure 4 shows the principle of five-base DNA methylation sequencing. As shown in Figure 4A, genomic DNA is fragmented into pieces up to several kilobases in length, which serve as the DNA template. Figure 4B illustrates DNA glycosylase excising a 5-MeC from the template (black), leaving an abasic site. During SMRT™ sequencing, the DNA polymerase synthesizes the complementary strand and preferentially incorporates a fluorophore-phospholinked pyrene analog opposite the abasic site. This fluorophore has spectral characteristics distinct from those of the other four labeled nucleotides and indicates the presence of a 5-MeC in the original template. Further, error metrics can also be used to identify the modification, e.g., an increase in binding events for the pyrene analog may occur at the abasic site, as well as at downstream positions as the incorporated pyrene analog is "buried" in the nascent strand during subsequent incorporation events. In certain embodiments, a non-hydrolyzable pyrene analog carrying a detectable label is used at a concentration sufficient to bind (and be detected) several times at the abasic site before a hydrolyzable (and, preferably, distinctly labeled) analog is incorporated. Methods using non-hydrolyzable analogs are further described below.

[0088] A potential challenge in carrying out the above-described methods is that many DNA glycosylases display some lyase activity, e.g., bifunctional DNA glycosylase/AP lyases. These enzymes can cleave the phosphodiester backbone 3' to the AP (abasic) site generated by the glycosylase activity resulting in an abasic and unsaturated ribose derivative at that site, which could prevent a polymerase from incorporating the pyrene analog complementary to this site. In certain cases, it may be desirable to suppress any lyase activity of the *Arabidopsis* repair enzymes and enhance the desired glycosylase activity. Strategies for achieving this include site directed mutagenesis and the addition of a catalytically inactive AP endonuclease to the glycosylase reaction. (See, e.g., A. E. Vidal, et al., *Nucleic Acids Res* **2001**, *29*, 1285,

incorporated herein by reference in its entirety for all purposes.) A parallel protein mutagenesis program aims to enhance polymerase processivity in the presence of a dPTP analog. Other variations exploit ways in which the kinetics of pyrene incorporation into the abasic site are effected by fluorophore identity, the number of phosphates attached to the pyrene analog, and the structure of the linker connecting the fluorophore to the terminal phosphate group.

[0089] In other embodiments of single-molecule, five-color DNA methylation sequencing, DNA glycosylase activity can be combined with addition of a non-natural base to replace the methylated base. Briefly, after glycosylase-catalyzed excision of 5-MeC (with or without cleavage of the phosphodiester backbone), a class I or class II AP endonuclease is added to remove the abasic ribose derivative by cleavage at the phosphate groups 3' and 5' to the abasic site, thereby leaving 3'-OH and 5'-phosphate termini. A polymerase capable of extending from the free 3'-OH (e.g., Pol I or human pol β) and a non-natural base (e.g., isoC, isoG, or ^{Me}isoC) are added to incorporate the non-natural base into the abasic site. A DNA ligase (e.g., LigIII) is added to close the phosphodiester backbone by forming covalent phosphodiester bonds between the free 3'-OH and 5'-phosphates via ATP hydrolysis. Finally, a processive polymerase (e.g., Φ29 DNA polymerase) is used to synthesize a nascent nucleic acid strand complementary to the template strand, where the fifth nucleotide analog is the complement of the non-natural base that replaced 5-MeC in the template. For example, if the replacement base was isoC or ^{Me}isoC, then the fifth analog would be isoG. As such, the fifth analog would only incorporate into the nascent strand at positions complementary to 5-MeC sites in the template nucleic acid. In preferred embodiments, the fifth analog has a detectable label (e.g., fluorescent dye) that is distinct from labels on other reaction components, e.g, detectable labels on other nucleotide analogs in the reaction mixture.

[0090] Further, glycosylases exist or can be engineered for various DNA modifications, damage, or mismatches, so the methods described above are applicable not only for detection of 5-MeC, but also provide methods for detecting those other types of modifications, as well. Methods for the use of glycosylases for detection of other types of DNA damage are described in U.S.S.N. 61/186,661, filed June 12, 2009 and incorporated herein by reference in its entirety for all purposes. In certain embodiments, the pyrene (or similar) nucleotide analog can be non-hydrolyzable to increase the residence time and, therefore, lengthen the emitted signal indicative of the presence of the particular lesion of interest. A non-hydrolyzable fifth-base is eventually displaced by a hydrolysable analog and synthesis of the nascent strand continues. Alternatively, a fifth-base may be hydrolysable but may produce multiple separate signals prior to incorporation to increase the likelihood of detection.

Bisulfite Modification

[0091] In certain embodiments, the template may be modified by treatment with bisulfite. Bisulfite sequencing is a common method for analyzing CpG methylation patterns in DNA. Bisulfite treatment deaminates unmethylated cytosine in a single-stranded nucleic acid to form uracil (P. W. Laird, *Nat Rev Cancer* **2003**, *3*(4), 253; and H. Hayatsu, *Mutation Research* **2008**, *659*, 77, incorporated herein by reference in their entireties for all purposes). In contrast, the modified 5-MeC base is resistant to treatment with bisulfite. As such, pretreatment of template DNA with bisulfite will convert cytosines to uracils, and subsequent sequencing reads will contain guanine incorporations opposite 5-MeC nucleotides in the template and adenine incorporations opposite the uracil (previously unmethylated cytosine) nucleotides. If a nucleic acid to be treated with bisulfite is double-stranded, it is denatured prior to treatment. In conventional methods, amplification, e.g., PCR, typically precedes sequencing, which amplifies the modified nucleic acid, but does not preserve information about the complementary strand. In contrast, certain embodiments of the present invention include use of a template molecule comprising both strands of a double-stranded nucleic acid that can be converted to a single-stranded molecule, e.g., by adjusting pH, temperature, etc. Treatment of the single-stranded molecule with bisulfite is followed by single-molecule sequencing, and because the template retains both strands of the original nucleic acid, sequence information from both is generated. Comparison of the resulting sequence reads for each strand of the double-stranded nucleic acid will identify positions at which an unmethylated cytosine was converted to uracil in the original templates since the reads from the two templates will be non-complementary at that position (A-C mismatch). Likewise, reads from the two templates will be complementary at a cytosine position (G-C match) where the cytosine position was methylated in the original template. In certain preferred embodiments, a circular template is used, preferably having regions of internal complementarity that can hybridize to form a double-stranded region, e.g., as described in U.S. 12/383,855 and U.S. 12/413,258, both filed on March 27, 2009, and both incorporated herein by reference in their entireties for all purposes.

[0092] As described elsewhere herein, methylcytosine has an effect on IPD over a number of neighboring positions when compared to non-methylated cytosine. Uracil compared to thymine is like unmethylated cytosine compared to methylcytosine (i.e. the only difference between U and T is that T has an additional methyl group). Thus, the invention provides methods for performing bisulfite sequencing in which the polymerase kinetics (IPD and pulse width) or the mismatch incorporation rate are monitored in addition to the actual nucleotides being incorporated. Detection of a change in either of these kinetic parameters or in the mismatch rate

at the position in question, or at neighboring positions, is used to determine whether or not a position was always a T or is a U that was originally an unmethylated cytosine.

[0093] In certain embodiments, polymerase mutants are designed that are more sensitive to the difference between thymine and uracil in order to enhance the effect described above. Methods for designing polymerase variants are described in detail above and need not be repeated here.

[0094] Additionally or alternatively, PCR of uracil-containing oligonucleotides is not necessarily as efficient as PCR without uracil. This issue can bias the PCR amplification of bisulfite-converted DNA. The methods of sequencing-by-synthesis using bisulfite-modified templates described herein circumvent this problem by not using PCR amplification. However, the kinetics of these sequencing-by-synthesis reactions can be monitored to detect changes in kinetics due to the presence of uracil residues.

[0095] Further, the methods presented herein are useful for detecting PCR bias in the amplification of bisulfite-treated nucleic acids. For example, a few rounds of PCR could be performed on various oligos, some with uracil and some without (including controls with the same sequence but containing thymine in place of uracil). After performing sequencing-by-synthesis on all the resulting oligos, one could determine the percentage of oligos that still contain uracil. If it's different than the expected percentage given ideal (unbiased) PCR amplification, then a bias has been detected.

[0096] In yet further embodiments, a template nucleic acid is exposed to a reagent that transforms a modified nucleotide to a different nucleotide structure. For example, a bacterial cytosine methyl transferase converts 5-MeC to thymine (M. J. Yebra, et al., *Biochemistry* **1995**, *34*(45), 14752, incorporated herein by reference in its entirety for all purposes). Alternatively, the reagent may convert a methyl-cytosine to 5-hydroxy-methylcytosine, e.g., TET1 (M. Tahiliani, et al., *Science* **2009**, *324*(5929), 930, incorporated herein by reference in its entirety for all purposes). In further embodiments, the reagent may include a cytidine deaminase that converts methyl-cytosine to thymine (H. D. Morgan, et al., *J Biological Chem* **2004**, *279*, 52353, incorporated herein by reference in its entirety for all purposes). In yet further embodiments, a restriction enzyme that specifically alters a modification of interest can be used to create a lesion at the modification site. For example, DPNI cleaves at a recognition site comprising methyladenosine. Optionally, the cleaved template could be repaired during an analytical reaction by inclusion of a ligase enzyme in the reaction mixture. As noted elsewhere herein, nucleotides other than 5-MeC can also be modified and detected by the methods provided herein. For example, adenine can be converted to inosine through deamination, and this conversion affected by methylation of adenine, allowing differential treatment and detection of adenine and MeA.

DMS Modification

[0097] In certain embodiments, the template may be modified by treatment with dimethyl sulfate (DMS) prior to sequencing. DMS is a chemical that methylates the N7 position of guanine in dsDNA, and to a lesser extent the N3 position of adenine in dsDNA. If proteins are bound to a DNA treated with DMS, the proteins will block the methylation of the sequences to which they are bound. The bound proteins can then be removed and the DNA treated with piperidine, which breaks the DNA backbone by removal of the methylated bases. Protected regions of the DNA are identified as having been bound to the proteins during the DMS treatment. DMS also modifies the N3 position of cytosine and the N1 position of adenine in single-stranded DNA or RNA so these bases can no longer base pair with their complement. Since both these positions are involved in base-pairing, regions that are double-stranded during DMS treatment are protected from modification. Reverse transcriptase PCR and gel analysis is subsequently used to identify regions that were unmodified, and are therefore likely regions that adopt secondary structures that protect them from DMS treatment.

[0098] The present invention provides methods for real-time, single-molecule sequencing of nucleic acids that have been subjected to DMS treatment as a means for detecting both binding sites of nucleic acid binding agents, as well as sites of secondary structure formation, e.g., G-quadruplex structures (also known as G-tetrads or G_4-DNA; see, e.g., Zheng, et al. (2009) "Molecular crowding creates an essential environment for the formation of stable G-quadruplexes in long double-stranded DNA," Nuc Ac Res 1-12, incorporated herein by reference in its entirety for all purposes). For example, dsDNA bound to one or more nucleic acid binding agents is subjected to DMS treatment, and the binding agents are subsequently removed. The resulting dsDNA is subjected to template-directed sequencing and pulse metrics are monitored to identify locations where guanine or adenine were methylated. For example, A and G template nucleotides that cause a distinguishable change in one or more pulse metrics are identified as not having been bound by the agent(s), and A and G template nucleotides that do not cause a distinguishable change in one or more pulse metrics are identified as having been bound by the agent(s). In certain embodiments, the DMS treatment takes place *in vivo*, and the dsDNA is subsequently extracted and sequenced to study transcription factor binding in the cell. Alternatively, dsDNA can be extracted from cells and subsequently exposed to one or more nucleic acid binding agents prior to treatment with DMS *in vitro*. The DMS treatment can performed in solution, or can be performed after the dsDNA is immobilized, e.g., at a reaction site. The nucleic acid binding agents that can be studied include, but are not limited to, transcription factors, polymerases, ribosomes, and associated cofactors to DNA, and thereby study which DNA regions are being actively transcribed in different cells, in healthy vs.

diseased tissue, in different cell cycle stages, in response to various environmental stimuli, and the like. For example, in certain embodiments DMS is applied *in vivo* or *in vitro* to mRNAs bound by actively translating or stalled ribosomes. The resulting mRNA templates are subsequently sequenced in real time, and the reactions are monitored for altered kinetics, which are indicative of modified bases. Alternatively, the DMS-treated mRNAs can be heated to degrade modified regions, leaving only unmodified regions for sequencing. The sequence data so generated is used to identify the mRNAs to which a ribosome was bound, and therefore the mRNAs that were being actively translated in the sample from which they were extracted. Other methods of ribosome profiling are known in the art, e.g., Ingolia, et al. (2009) Science 324(5924):218-23, the disclosure of which is incorporated herein by reference in its entirety for all purposes.

[0099] In further embodiments, DNA and RNA secondary structure profiling can be performed by applying DMS to single-stranded DNA or RNA (e.g., mRNA, siRNA, microRNA, rRNA, tRNA, snRNA, etc.) and sequencing the DMS-modified nucleic acid using an appropriate polymerase. (Methods for sequencing RNA molecules using RNA dependent polymerases are described in detail in U.S.S.N 61/186,661, filed June 12, 2009 and incorporated herein by reference in its entirety for all purposes.) Regions of the treated nucleic acid that elicit altered polymerase kinetics are identified as regions that were single-stranded during the DMS treatment, and regions of the treated nucleic acid that do not elicit altered polymerase kinetics are identified as regions that were double-stranded during the DMS treatment and therefore likely contained duplex secondary structure, e.g., hairpins. In certain embodiments, the nucleic acid is heated prior to sequencing to cause degradation of the modified regions. The remaining, undegraded nucleic acid is subsequently subjected to sequencing and the sequence data so generated is used to identify regions of the original nucleic acid that formed secondary structures that prevent DMS modification.

[00100] DMS modification can also be used to map regions that form non-B-form secondary structures, some of which have regulatory roles *in vivo*. For example, G-quadruplexes consist of stacks of Gs that protect the guanosines from DMS-modification, even in the absence of a nucleic acid binding agent. Subsequent sequence analysis is used to identify regions that were protected from DMS modification, and therefore are likely to have had some protective secondary structure.

[00101] Further, although described primarily in terms of DMS modification, other types of chemical and/or enzymatic modifications can also be used in an analogous fashion, as will be clear to one of ordinary skill in the art based on the teachings herein. For example, other methods of DNA or RNA footprinting are particularly useful in the methods herein, including,

e.g., use of DNaseI, hydroxyl radicals, or UV irradiation for cleavage of nucleic acid that is not bound by an agent. Such methods are described more fully in the published literature.

[00102] The template altered by exposure to the reagent is sequenced, e.g., using a real-time, single-molecule methodology such as SMRT™ sequencing. In certain preferred embodiments, the sequencing is performed multiple times on the same template, e.g., by rolling-circle synthesis or another form of molecular redundant sequencing. The loci in the template containing altered nucleotides are identified by analysis of the resulting sequence reads. In cases in which the 5-MeC nucleotides were converted to non-altered nucleotide (e.g., thymine), molecular redundant sequencing on both the forward and reverse strands is useful for further refining the identification of the altered nucleotides since the transformation disrupts the normal Watson-Crick base pairing. For example, if MeC•G pair is converted to T•G, the forward and reverse reads will show non-complementary nucleotides at that position (A and C), indicating that the base pair in the template was non-standard, likely due to an alteration of a 5-MeC at that position. Methods for molecular redundant sequencing are further described in U.S. Patent No. 7,476,503 and U.S. Application Serial Nos. 12/383,855 (filed March 27, 2009), 12/413,258 (filed March 27, 2009), 12/413,226 (filed March 27, 2009), and 12/561,221 (filed September 16, 2009), all of which are incorporated herein by reference in their entireties for all purposes.

VI. Detection of Agent-Nucleic Acid Interactions

[00103] Another example of a biological process that may be monitored in accordance with the invention is association of a nucleic acid binding agent (e.g., a protein, nucleic acid, or small molecule) with a single nucleic acid molecule. As for the chemical modifications to the template described above, use of such agents serve to enhance the detection of the modification, e.g., by amplifying a signal indicative of the modification. Many types of agents bind to nucleic acids, such as transcription factors, RNA and DNA polymerases, reverse transcriptases, histones, nucleases, restriction enzymes, replication protein A (RPA), single-stranded binding protein (SSB), anti-DNA antibodies, DNA damage-binding agents, agents that bind altered nucleotides (e.g., methylated), small RNAs, microRNAs, drug targets, etc. In particular, transcription factors are involved in gene expression regulation and are thus very important for the study of diseases such as cancer. Further, RPA binds single-stranded DNA during replication to keep DNA unwound and accessible to the polymerase. Current technologies for detecting the binding of a protein transcription factor to a DNA molecule involve bulk detection. Certain aspects of the invention provide a method for detecting the binding of a transcription factor or other nucleic acid binding agent to a single molecule of DNA. The advantages of the methods described herein include, but are not limited to, improved resolution of kinetics (e.g., of

association and dissociation), binding loci, and statistical analysis; and greater sensitivity and simplicity.

[00104] In certain aspects, the invention provides detection of binding of a nucleic acid binding agent onto a single nucleic acid molecule through a technology that involves observing the activities of single molecules of polymerases in real time and with high multiplex capabilities, thereby allowing the screening of multiple nucleic acid binding agents (or other components of the reaction) with high throughput. In particular, the invention employs analogous processes used for single-molecule, real-time DNA sequencing, and with some modifications, exploits such processes to characterize various aspects of binding of nucleic acids by proteins of interest. Such sequencing technology has been previously described, e.g., in Eid, et al. (incorporated herein above). In certain preferred embodiments, one or more components of the reaction are immobilized at a reaction site, e.g., in an optical confinement such as a ZMW. Alternatively or additionally, multiple reactions can be simultaneously monitored by immobilizing them at discrete locations on a substrate, e.g. in an array of optical confinements. Further, to prevent displacement of the agent prior to a detectable affect on the reaction (e.g., a pause), the binding may be enhanced through various alterations to the reaction mixture (e.g., salt concentration, pH, temperature, etc.), or through alterations to the agent itself. For example, a DNA-binding protein may comprise various mutations that enhance binding under the conditions of the sequencing reaction, e.g., by lowering the Kd of the binding domain (e.g., a methyl binding domain) or by duplicating the domain to increase the effective concentration of the binding domain in the vicinity of the DNA template.

[00105] In certain preferred embodiments, a single nucleic acid template is bound to a sequencing engine (e.g., a polymerase or reverse transcriptase) that is synthesizing a nascent nucleic acid strand, e.g., during a template-directed sequencing reaction or a sequencing-by-synthesis reaction. The template can be any nucleic acid template appropriate for template-directed sequencing, e.g. single-stranded or double-stranded DNA, RNA, or a DNA/RNA hybrid. Further, the nucleic acid template can be linear or circular. For example, a dsDNA template can be bound by a polymerase in an optical confinement, e.g., a ZMW, as described above and in, e.g., Foquet, et al., and Levene, et al., both of which are incorporated herein supra. A nucleic acid binding agent, such as a transcription factor or DNA damage-binding agent, is added to the reaction mixture under conditions that promote binding of the agent to the template. If the template is bound by the agent in a location ahead of the polymerase, the bound agent impedes the translocation of the polymerase along the template, resulting in a pause or full stop in polymerization at or adjacent to the position at which the agent bound. Real-time monitoring of the ongoing sequencing reaction will allow detection of the pause or stop, which is indicative

of (i) the fact that the agent bound the template, and (ii) the position on the template that was bound by the agent, e.g., based on the sequence of nucleotides incorporated immediately prior to the pause or stop. Further, a consensus sequence for the binding site of the agent can be determined by statistical analysis of the "binding-affected" (e.g., containing a pause or truncated) sequence reads generated in the presence of the agent and the non-binding-affected (e.g., full-length) sequence reads generated in the absence of the agent. For example, truncated sequence reads (or sequence reads having detectable pauses) generated in the presence of the agent provide a location on the template at which the polymerase was blocked, and full-length reads generated in the absence of the agent provide the binding site sequence. In certain embodiments, sequence reads from the region of the template immediately downstream of the point at which the polymerase is blocked are analyzed together to find a sequence (specific or degenerate) they have in common, and this common sequence is identified as the consensus binding site for the agent. Such analyses are routine in nucleic acid sequence analysis and require no further elaboration here.

[00106] In certain embodiments, a nucleic acid binding protein of interest is introduced into a reaction mixture comprising a pool of nucleic acid templates. The pool of templates is exposed to the protein under conditions that promote binding, and polymerase enzymes are subsequently added to the reaction mixture and allowed to bind the templates, e.g., at a single-stranded region comprising a bound oligonucleotide primer. The reaction mixture further comprises a set of detectably labeled nucleotides, wherein each type of nucleotide in the set is linked to a distinct label that is optically identifiable during polymerization, thereby providing a distinct signal for each nucleotide incorporation event that identifies the base incorporated into the nascent strand. The polymerase-template complexes are immobilized on a substrate such that signals emitted from each complex are optically resolvable from signals emitted from every other complex on the substrate. Preferably, the reaction mixture is lacking a component required for polymerization to prevent polymerase activity prior to immobilization. Such a component is subsequently added to the reaction mixture allowing the polymerase to commence synthesis of a nucleic acid strand complementary to the template to which it is bound. For those templates that were not bound by the protein, synthesis continues unimpeded and the template is fully sequenced in the optical confinement, generating a full-length sequence read for the template. In contrast, the templates that were bound by the protein are processed by the polymerase until the bound protein is encountered on the template, at which time the polymerase will pause or stop polymerizing the complementary strand. The truncated sequence read generated from such a stalled polymerase-template complex will provide sequence information for the template upstream of the protein binding site. Statistical analysis of this sequence information, both at the

single molecule level and across the pool of templates, can be used to both identify the particular nucleic acid templates bound (or not bound) by the protein, as well as identifying the position at which the protein binds. For example, this technique can be used to map specific protein binding sites on the template, e.g., sequence-specific or lesion/damage-specific binding sites.

[00107] This assay can by easily modified to test the impact of various reaction conditions, e.g., pH, temperature, ion concentrations, and presence or absence of agents such as drugs, antibodies, or binding competitors. These tests can be used to identify optimal reaction conditions, e.g., for causing a pause or stop in an ongoing sequencing reaction or for binding to a particular subset of the pool of template nucleic acids. Further, the assay can be used to test variants and/or mutants of known nucleic acid binding proteins to screen such mutants for desired characteristics, such as binding under stringent conditions or having altered sequence specificity for binding. The assay can also be used to test variants and/or mutants of polymerase enzymes for desired characteristics, such as the ability to bypass a particular nucleic acid binding protein. Further, the specificity of binding can be explored by performing the assay with different pools of nucleic acid templates.

[00108] In certain embodiments, the nucleic acid binding protein is a transcription factor (TF) with a specific consensus binding sequence, e.g., TGACTCA for AP1 or GGACTTCC for NF-κB. DNA template molecules that contain the consensus binding sequence are bound by the TF at that sequence, and those that do not are not bound by the TF. When the translocating polymerase encounters a bound TF, the polymerase stops polymerizing, and the cessation of signals emitted from the complex is indicative that the TF bound the template and, therefore, that the template contains the consensus binding sequence. As noted above, various reaction conditions can be tested for their effect on either the binding of the TF or the ability of the polymerase to bypass it or displace it from the template. Statistical analysis of the sequence information from the DNA templates that were bound by the TF can be used to further characterize the TF, e.g., by (i) identifying genes targeted by the TF, e.g., using publicly available genome sequence data; (ii) identifying the consensus binding sequence, e.g., using sequence data generated from the same templates in the absence of the TF; (iii) studying the interaction of multiple transcription factors; (iv) modulation of TF binding by other proteins, small molecules, etc.; (v) testing the temperature sensitivity of binding; (vi) identifying and characterizing the abundance of particular DNA-binding proteins, e.g., in a cell extract; and the like. For example, the identity and abundance of DNA-binding proteins can be compared between a) different tissues, cell lines, cell developmental stages, species, or subspecies; b) healthy and diseased samples; and c) in the presence and absence of environmental stressors and/or various agents (e.g., drugs, toxins, etc.). Yet further, variants and mutants of different

components of the reaction mixture, e.g., TF, polymerase, template, etc., can be tested to identify those with particularly desirable characteristics, e.g., tight binding, protein displacement activity, non-consensus binding sequences with higher binding affinity to the TF, etc.

[00109] When implemented on in an arrayed format, such investigations would be highly parallel, enabling high-throughput screening assays. Arrays of reactions are carried out on highly multiplexed confocal fluorescence microscope systems (see, e.g., Lundquist, et al., incorporated herein above) in which the instrument detects fluorescent signals from each reaction site on the array, resulting in a highly parallel operation. Although preferred embodiments use arrays of zero mode waveguides, as described elsewhere herein, these assays could also be performed in other systems capable of real-time single-molecule detection, e.g., using total internal reflection fluorescence (TIRF) microscopy or waveguide technology.

[00110] Although certain embodiments are described in terms of nucleic acid binding proteins, it will be appreciated that the methods and systems described herein are equally applicable to other nucleic acid binding agents capable of pausing, stopping, or otherwise disrupting processive template-directed synthesis of a nascent nucleic acid molecule, e.g., nucleic acids and analogs and mimetics thereof (e.g., protein nucleic acids), lipids, sugar-oligoamides, intercalating dyes, major and minor groove binders, etc.

VII. Nucleic Acid Binding Agents as Analytical Tools

[00111] In certain aspects, nucleic acid binding agents are used in the methods, compositions, and systems of the invention to detect and/or reverse modifications in nucleic acid molecules. Such agents are typically used to enhance the response of a polymerase to a modification in the template nucleic acid. That is, the methods herein can be used to detect binding of an agent to the template, whether in response to a modification as described below, or simply an unmodified recognition site within the sequence of the template, as described above. Further, the effects of various agents on the creation, detection, or bypass of a nucleotide modification can also be tested and compared. For example, a template can be treated in various different ways (e.g., with and without a nucleic acid binding agent) and subsequently subjected to single-molecule sequencing-by-synthesis, which is monitored for a disruption in sequence read generation that is characteristic of binding of the agent to the template. In other embodiments, a template containing a known modification can be subjected to single-molecule sequencing-by-synthesis in the presence of various agents or reaction conditions. The reaction is monitored for the activity of the polymerase on the modified template to determine if the presence of any of the agents or other conditions impacts the ability of the polymerase to bypass or pause at the modification.

[00112] In certain specific embodiments, accentuating the differences in interpulse duration and/or pulse width between methylated and unmethylated DNA involves DNA binding proteins. It has been shown that some DNA polymerases stall when they encounter a DNA-bound protein complex. (See, e.g., M. Elias-Arnanz, et al., *EMBO J* **1997**, *16*, 5775, incorporated herein by reference in its entirety for all purposes.) In SMRT™ sequencing, this stall is detected as an unusually long interpulse duration that would end when the binding protein dissociates from the DNA template or is displaced by the translocating polymerase. There are a number of proteins that can bind stably and specifically to methylated DNA including members of the MBD family of human proteins, all of which contain a methyl-CpG binding domain (MBD). For example, MECP2, MBD1, MBD2, and MBD4 all bind specifically to methylated DNA, and are involved in repressing transcription from methylated gene promoters. Binding of these proteins to a template nucleic acid is expected to cause a translocating polymerase to pause proximal to the bound protein. As such, an increased pause duration during single-molecule sequencing reactions is indicative of a methylated base in the template nucleic acid. It is therefore important that the protein bind tightly to its target nucleic acid sequence. Natural MBD proteins only have micromolar Kd affinities for methyl-CpG sequences, so engineered MBD proteins that bind more tightly to the methylated template sequence can enhance detectability of methylated bases. For example, a multimerized MBD1 protein is provided in Jorgensen, et al., *Nucleic Acids Research* **2006**, 34(13), e96. Such engineered proteins can have a single methyl binding domain with a lower Kd (sub-micromolar) or multiple methyl-binding domains that increase the effective concentration of the methyl-binding domain in the vicinity of the methylated DNA template. More information on the MBD family of proteins is provided, e.g., in B. Hendrich, et al., *Mol Cell Biol* **1998**, 18(11), 6538; and I. Ohki, et al., *EMBO J* **2000**, 18(23), 6653.

[00113] In addition, the mammalian UHRF1 (ubiquitin-like, containing PHD and RING finger domains 1) protein binds tightly to methylated DNA and is required for its maintenance. Crystal structures of the SRA domain of this protein bound to DNA show that the 5-MeC is flipped out of the DNA duplex and stabilized by hydrophobic stacking and hydrogen bonding to SRA protein residues. (See, e.g., G. V. Avvakumov, et al. and H. Hashimoto, et al., both *supra.*) Finally, the monoclonal antibody to 5-MeC, used for methylated DNA immunoprecipitation, also binds specifically to methylated cytosine. (See, e.g., N. Rougier, et al., *Genes Dev* **1998**, *12*, 2108 ; and M. Weber, et al., *supra*, which are incorporated herein by reference in their entireties for all purposes.) All of the above-mentioned proteins are candidates for interfering with normal DNA polymerase processivity during SMRT™ sequencing. In order to enhance their polymerase stalling effects, these proteins can also be engineered to increase their affinity for

methylated DNA sites. See, e.g., H. F. Jorgensen, et al., *Nucleic Acids Res* **2006**, *34*, e96, the disclosure of which is incorporated by reference herein in its entirety for all purposes.

[00114] In yet further embodiments, an antibody against 5-MeC could be used to bind 5-MeC in a template nucleic acid, similar to the process used in methylated DNA immunoprecipitation assays (M. Weber, et al., *Nat Genet* **2005**, 37, 853). As such, the antibody essentially acts as an enhancer of the signal indicating the presence of the modification in the template by virtue of altering the polymerase dynamics. Various components of such reactions can be detectably labeled, e.g., the antibody, template, incorporated nucleotides, and combinations thereof, as described further elsewhere herein.

[00115] In still further embodiments, methyltransferases can be used to further facilitate detection of methyl-modified template nucleic acids. As described above, DNA methyltransferases catalyze the addition of methyl groups to DNA based upon recognition of methylation sites. For some methyltransferases (e.g., maintenance methyltransferases), the most active binding site in a nucleic acid is a hemi-methylated site in which one strand of the nucleic acid is methylated and the opposite strand is not. An enzymatically inactive methyl transferase (i.e., one that is unable to methylate nucleic acids) will therefore preferably bind to a hemi-methylated strand of DNA. In a real-time, template-directed sequencing reaction, a methylated single-stranded template becomes hemimethylated after nascent strand synthesis. A detectably labeled methyltransferase can therefore be detected interacting with the hemimethylated product of the synthesis reaction in real-time.

[00116] In certain embodiments, a circular template is used to permit rolling-circle synthesis by the polymerase in which a single-stranded circular template is converted to a double-stranded circular template. In preferred embodiments, the polymerase is capable of strand displacement such that after proceeding around the template once it begins to displace the nascent strand ahead of it as synthesis continues. This process eventually results in long concatemer containing multiple copies of the complement to the original template molecule. In such a system, a single-stranded methylated template is converted to a double-stranded hemimethylated template. A methyltransferase present in the reaction mixture can bind the hemimethylated sites and, if detectably labeled, this binding can be readily monitored in real time. When the polymerase encounters a bound methyltransferase, a pause may be detected prior to dissociation of the methyltransferase. The location of the pause in the resulting sequence reads can be used to map the position of the methylated site within the template molecule, even in the absence of a detectable label on the methyltransferase. For example, the pause can be used to identify the binding of the methyltransferase, e.g., in cases in which the methyltransferase is not detectably labeled, and in such cases the methyltransferase would essentially serve to extend

the pause at a methylated site, thereby facilitating identification of such a site in the template nucleic acid.

[00117] Other types of modifications can also be detected and/or reversed by nucleic acid binding agents. For example, among all types of DNA damage, oxidative base damage by reactive oxygen species (ROS) has been recognized as a major cause of cell death and mutagenesis in aerobic organisms (see, e.g., Finkel, et al. (2000) Nature 408(6809): 239-47, which is incorporated herein by reference in its entirety for all purposes). DNA oxidative lesions are primarily recognized and repaired by base excision repair (BER) pathways (see, e.g., Fromme et al. (2004) Adv Protein Chem 69: 1-41, which is incorporated herein by reference in its entirety for all purposes). In humans, the BER pathway for detecting and repairing a common oxidative lesion, 7,8-dihydro-8-oxoguanine ("8-oxoG"), begins with recognition of the lesion by a human oxoguanine DNA glycosylase 1 (hOgg1), which is a DNA glycosylase/apurinic (AP) lyase (see, e.g., Klungland, et al. (2007) DNA Repair (Amst) 6(4): 481-8, which is incorporated herein by reference in its entirety for all purposes).

[00118] The modified base 8-oxoG is discussed at length *supra*. Recent fluorescence and crystallography studies of hOgg1 found that this DNA glycosylase recognizes the oxidative DNA lesion 8-oxoG by scanning the DNA duplex, flipping the DNA base out, and transferring the damaged base from a pre-sampling binding site to the damage recognition binding site. Single-molecule experiments revealed the rapid sliding activity of hOgg1 on DNA duplex. For more detailed information on these studies, see Banerjee, et al. (2005) Nature 434(7033): 612-8, and Blainey, et al. (2006) Proc Natl Acad Sci USA 103(15): 5752-7, the disclosures of which are incorporated herein by reference in their entireties for all purposes.

[00119] In certain embodiments, the methods provided by the invention expose a nucleic acid template to a damage-recognition agent that binds to the template at a damaged nucleotide in a manner that blocks bypass of the lesion by a polymerase translocating along the template. The blockage causes a cessation of incorporation-dependent signaling from the reaction site, thereby indicating the damage-recognition agent has bound a damaged nucleotide in the template. In some aspects, the methods further include exposing a damaged template to additional reaction components that act to repair the damage, restoring the template and allowing dissociation of the damage-recognition agent from the previously damaged nucleotide. Elements of the damage-repair (e.g., base excision repair (BER)) machinery can be provided in the original reaction mixture, or can be added to an ongoing reaction. If the polymerase pauses but does not dissociate, the polymerization reaction can continue after DNA repair has been completed and the repair machinery has dissociated from the template or translocated away from the previously-damaged site.

[00120] In preferred embodiments, the damage-recognition agent is a protein involved in BER such as DNA gycosylases/apurinic (AP) lyases, e.g., hOGG1 (human oxoguanine DNA glycosylase 1), yOGG1 (yeast homolog of hOGG1), FPG protein (MutM; bacterial homolog of hOGG1); and others known in the art. Other proteins that can be used as a damage-recognition agent include other DNA glycosylases, e.g., AlkA, Nth, Nei, MutY, uracil DNA glycosylases (UDG), single-strand selective monofunctional uracil-DNA glycosylase (SMUG), thymine DNA glycosylase (TDG), NEIL (e.g., hNEIL1 and hNEIL2), etc. Reaction components for repair of a damaged template bound by the damage-recognition agent include, e.g., AP endonucleases, DNA polymerase beta, and ligase, among others known in the art. See, e.g., McCullough, et al. (1999) Annu Rev Biochem 68: 255-85, which is incorporated herein by reference in its entirety for all purposes. Further, additional proteins that stimulate damage recognition may also be included in an analytical reaction; e.g., HAP1 (APE1) protein has been found to stimulate hOGG1 activity (Vidal, et al. (2001) Nuc. Ac. Res. 29(6):1285-1292).

[00121] In certain embodiments, more than one polymerase may be present in a template-directed sequencing reaction in which one or more lesions may be present on the template nucleic acid. For example, "bypass polymerases" have been discovered in both prokaryotes and eukaryotes, most of which belong to the Y-family of polymerases and/or are considered to be repair polymerases. In contrast to replicative polymerases, they operate at low speed, low fidelity, and low processivity. However, because their active sites adopt a more open configuration than replicative polymerases they are less stringent and can accommodate altered bases in their active sites. For more information on bypass polymerases, see, e.g., Cordonnier, et al. (1999) Mol Cell Biol 19(3):2206-11; Friedberg, et al. (2005) Nat Rev Mol Cell Biol 6(12):943-53; Holmquist, et al. (2002) Mutat Res 510(1-2):1-7; Lehmann, A.R. (2002) Mutat Res 509(1-2):23-34; Lehmann, A.R. (2006) Exp Cell Res 312(14):2673-6; Masutani, et al. (1999) Nature 399(6737):700-4; and Ohmori, et al. (2001) Mol Cell 8(1):7-8, the disclosures of which are incorporated herein by reference in their entireties for all purposes. Certain of these polymerases can bypass lesions in a nucleic acid template and carry out "translesion synthesis" or TLS. As such, DNA replication in the presence of such lesions was found to require multiple polymerases and the "polymerase switch model" was developed (see, e.g., Friedberg, et al. (2005) Nat Rev Mol Cell Biol 6(12):943-53; Kannouche, et al. (2004) Cell Cycle 3(8):1011-3; Kannouche, et al. (2004) Mol Cell 14(4):491-500; and Lehmann, et al. (2007) DNA Repair (Amst) 6(7):891-9, all of which are incorporated herein by reference in their entireties for all purposes). In brief, the polymerase switch model is model for lesion bypass during replication that involves replacement of a replicative polymerase with a bypass polymerase at a lesion, synthesis of the nascent strand by the bypass polymerase until past the lesion, and subsequent

replacement of the bypass polymerase with the more processive, higher fidelity replicative polymerase for continued replication past the lesion.

[00122] In certain preferred embodiments, one or more bypass polymerases is included in a template-directed nucleic acid sequencing reaction. For example, during the course of a reaction in which a replicative polymerase encounters and is blocked by a lesion in a template nucleic acid, the replicative polymerase is replaced by a bypass polymerase at the site of the lesion, and the bypass polymerase synthesizes a segment of the nascent strand that is capable of base-pairing with the damaged base, and may further include one or more bases prior to and/or past the site of the lesion in a process called "translesion synthesis." The limited processivity of the bypass polymerase causes it to dissociate and be replaced by the replicative polymerase following translesion synthesis. The replicative polymerase continues to synthesize the nascent strand until another blocking lesion is encountered in the template, at which point it is once again replaced by a bypass polymerase for translesion synthesis. (See, e.g., Friedberg, et al. (2005) Nat Rev Mol Cell Biol 6(12):943-53; and Kannouche, et al. (2004) Mol Cell 14(4):491-500, incorporated herein by reference above.) The process continues until the template has been replicated or the reaction is terminated, e.g., by the investigatior. One particular advantage of the polymerase switch method of template-dependent sequencing is that is it tolerant of most types of lesions in the template nucleic acid. As such the damaged template can be sequenced through a lesion, thereby allowing reinitiation of synthesis downstream of the lesion and increasing read lengths on lesion-containing templates.

[00123] Various different bypass polymerases known to those of ordinary skill in the art can be used with the methods and compositions provided herein, include prokaryotic polymerases (e.g., DNA polymerase IV, polymerase V, Dpo4, Dbh, and UmuC) and eukaryotic polymerases (e.g., DNA polymerase η, DNA polymerase ι, DNA polymerase κ, and Rev1). In eukaryotes, multiple bypass polymerases participate in translesion synthesis, and a processivity factor, proliferating cell nuclear antigen ("PCNA"), is also required and can be included in a sequencing reaction.

[00124] In certain preferred embodiments, the template or primer is immobilized during the template-dependent synthesis reaction to ensure that the template remains at the reaction site during polymerase switching. Alternatively or additionally, one or more polymerases can be immobilized at the reaction site. Various immobilization strategies useful in different aspects of the invention are provided elsewhere herein.

[00125] Since the portion of the nascent strand corresponding to the site of the lesion in the template is synthesized by a bypass polymerase, the sequence reads generated therefrom are expected to be less reliable than those generated from regions of the nascent strand synthesized

by the replicative polymerase. As such, generation of redundant sequence information during a sequencing reaction is a preferred means of generating complete and accurate sequence reads. Redundancy can be achieved in various ways described elsewhere herein, including carrying out multiple sequencing reactions using the same original template with the sequence data generated in the multiple reactions combined and subjected to statistical analysis to determine a consensus sequence for the template. For example, the sequence data from a region in a first copy of the template that was replicated by a lower fidelity bypass polymerase can be supplemented and/or corrected with sequence data from the same region in a second copy of the template that was replicated with a higher fidelity replicative polymerase. Further, a template can be amplified (e.g., via rolling circle amplification) to generate a concatemer comprising multiple copies of the template that is subsequently sequenced to generate, a sequencing read that is internally redundant. The sequence data from a first segment of the concatemer (corresponding to a first region of the template) that was replicated by the bypass polymerase can be supplemented and/or corrected with sequence data from a second segment of the concatemer (that also corresponds to the first region of the template) that was replicated by the replicative polymerase. Further, as noted above, redundancy can also benefit identification and characterization of lesions that occur in the same position in a plurality of templates, or that occur at a single position in a template that is subjected to resequencing. For example, since base incorporation by the bypass polymerase is promiscuous, replicate sequencing reads for the region containing the lesion may show more than one "complementary base" being incorporated at the same position in different reads (of the same or an identical template), and detection of such promiscuity is indicative that there is a lesion at that position in the template nucleic acid(s).

[00126] In certain embodiments, a polymerase in the reaction mixture may comprise a detectable label to indicate when that polymerase is associated with the template nucleic acid. For example, a bypass polymerase can comprise a detectable label that will indicate when the bypass polymerase is carrying out translesion synthesis. The nucleotides incorporated into the nascent strand during that time can therefore be identified and "tagged" as corresponding to a region of the template that contains one or more lesions, thereby allowing targeting of statistical analysis to these sequence reads, e.g., as described above.

[00127] In yet further embodiments, a nucleic acid binding agent specifically binds to secondary structure in the nucleic acid template, e.g., hairpin loops, stem-loops, internal loops, bulges, pseudoknots, base-triples, and the like. Binding of an agent to such structures inhibits passage of the polymerase through the structures to a greater extent than the enzyme is inhibited in the absence of the agent, thereby increasing the resulting pause time and facilitating detection of the secondary structure. Examples of agents that have binding specificity for specific

structures and/or strandedness in nucleic acids include, e.g., intercalating agents, nuclease-deficient endonucleases (e.g., with a specificity for a double-stranded region within a stem-loop structure), polymerases, and various eukaryotic initiator proteins.

[00128] As noted above, various different types of templates for template-directed polymerization reactions can be used, e.g., single-stranded or double-stranded DNA, single-stranded or double-stranded RNA, and analogs and mimetics thereof. Further, the template can contain a combination of single-stranded and double-stranded regions, e.g., such as the templates described in U.S.S.N. 12/383,855 and 12/413,258, both filed on March 27, 2009 and incorporated herein by reference in their entireties for all purposes. The type of template used is limited only by the substrate specificity of the polymerase and damage-binding agent in the reaction. For example, Figure 5 provides an illustrative embodiment of such a reaction comprising a linear template and a damage-binding agent that recognizes a lesion in a single-stranded template. In A, the damage-binding agent (305) is scanning a linear, single-stranded nucleic acid template (310) ahead of a polymerase (315) performing template-directed polymerization of a nascent nucleic acid strand (320). In B, the damage-binding agent (305) has detected and bound to a lesion (325) in the single-stranded template (310). In C, the polymerase (315) has caught up with the damage-binding agent (305) and its progress along the template (310) is blocked. In D, the lesion has been repaired by repair machinery (330) recruited by the damage-binding agent (305). In E, the repair machinery has dissociated from the template (310) and the damage-binding agent (305) has translocated away from the previously damaged site, thereby allowing the polymerase (315) to resume synthesis of the nascent strand (320).

[00129] In some embodiments, a damage-binding agent with specificity for double-stranded nucleic acid may be used in a reaction comprising a single-stranded template, e.g., when the scanning and damage detection/binding is expected to occur after the polymerase has converted the single-stranded template to a double-stranded template by template-dependent polymerization, e.g., after a single-stranded circle has been converted to a double-stranded circle during "rolling-circle replication." For example, although an initial substrate in a reaction is a circular single-stranded nucleic acid template, after a polymerase has processed the template one time it becomes a double-stranded template and an appropriate substrate for a damage-binding agent that specifically scans and binds double-stranded nucleic acid. For example, Figure 6 illustrates an embodiment comprising a circular template and a damage-binding agent that recognizes a lesion in a double-stranded template. In A, the damage-binding agent (405) is scanning a circular, double-stranded nucleic acid template (435) ahead of a polymerase (415) performing template-directed polymerization while displacing the 5' end of the nascent nucleic acid strand being synthesized (440). In B, the damage-binding agent (405) has detected and

bound to a lesion (425) in the double-stranded template (435), and the progress of the polymerase (415) is blocked by the bound damage-binding agent (405). In C, the lesion has been repaired and the damage-binding agent (405) has translocated away from the previously damaged site, thereby allowing the polymerase (415) to resume synthesis of the nascent strand (420).

[00130] Although various embodiments are described in terms of recognition and, optionally repair of 8-oxoG lesions, other types of DNA damage can also be addressed by the methods herein. For example, in the case of hOGG1, the *N*-glycosylase activity releases damaged purines from double-stranded DNA, generating an apurinic (AP) site. The AP-lyase activity cleaves 3′ to the AP site leaving a 5′ phosphate and a 3′-phospho-α, β-unsaturated aldehyde. In addition to 8-oxoG (when paired with cytosine), hOGG1 also recognizes and removes 8-oxoA (when base paired with cytosine), foramidopyrimidine (fapy)-guanine and methy-fapy-guanine (Bjoras, M. et al. (1997) *EMBO J.*, 16, 6314-6322; and Boiteux, S. and Radicella, J. (1999) *Biochimie*, 81, 59-67, the disclosures of which are incorporated herein by reference in their entireties for all purposes). Other types of DNA damage that can be bound and, optionally, repaired by the methods herein include BER enzymes that repair other DNA base lesions (small DNA base modification), e.g. AAG/MPG for methylated lesions, UDG/SUMG1 for repairing uracil in DNA, APE for abasic sites, etc. Also included are nucleotide excision repair (NER) enzymes that repair more bulky DNA lesions, such as DNA base adducts and DNA intra- and inter-strand crosslinks. Furthermore, although the DNA polymerase switch methods described above are suitable for detecting and bypassing most DNA lesions that block a replicative polymerase, certain small base modifications like 8-oxoG can be bypassed by a replicative polymerase, and thus methods that include binding agents that block the polymerase at the site of a lesion can help ensure that such lesions are detected, and optionally removed, from the template to prevent the sequence data generated from the template-dependent sequencing reactions to be adversely affected.

[00131] In certain embodiments, hOGG1 is included in a template-directed DNA sequencing reaction in the presence of a polymerase and a set of nucleotides, each of which bears a label that is optically detectable and that distinctively identifies the base (e.g., A, G, T, or C). Detection of an optical signal upon interaction with the polymerase and incorporation into the nascent strand allows the practitioner to identify the base incorporated and, by complementarity, the sequence of the template DNA molecule. In preferred embodiments, the incorporation of nucleotides into the nascent strand continues in a processive fashion, generating an ordered set of optical signals that can be analyzed to provide a sequence for both the nascent strand and, by complementarity, the template strand. The hOGG1 enzyme associates with the

template, "scans" for damage, and specifically binds to locations at which such damage occurs. As such, if the template DNA molecule contains or acquires (e.g. during the course of the analytical reaction) DNA damage recognized by hOGG1, it is bound by hOGG1, bypass of the lesion by the polymerase is blocked, and the incorporation-based signal is slowed or stopped (e.g., by stalling or dissociation of the polymerase). Although such a blockage can cause dissociation of the polymerase, in certain preferred embodiments the polymerase merely pauses until the damaged nucleotide is repaired and hOGG1 and any other repair machinery dissociates from the template, at which time polymerization resumes and additional sequence data is generated from the template at and downstream of the site of the previously damaged nucleotide.

[00132] In certain preferred embodiments, one or more reaction components is immobilized at a reaction site, e.g., in an optical confinement such as a zero mode waveguide (ZMW). In some embodiments, the polymerase is immobilized and the nucleic acid template and damage-binding agent are free in solution. Methods for immobilizing a polymerase enzyme are available in the art and provided elsewhere herein. In other embodiments, the nucleic acid template can be immobilized at the reaction site with the polymerase and damage-binding agent free in solution. For example, in preferred embodiments the damage-binding agent translocates upon the template faster than the polymerase so it does not impede progress of the template-dependent sequencing reaction on an undamaged template. However, upon binding a lesion, the damage-binding agent will stop and bind to the site, blocking progress of a translocating polymerase past the lesion. For example, hOGG1 translocates much faster than phi29 polymerase on undamaged DNA, but after encountering a damaged nucleotide the enzyme will bind to the site and wait for other components of the BER machinery. Alternatively or additionally, the damage-binding agent may be immobilized at the reaction site. For example, in the case of hOGG1 only a single enzyme is required for DNA binding, scanning, and lesion recognition. Immobilization of a single damage-binding agent at the reaction site increases the likelihood that a single template at each reaction site will be scanned for damage. Methods for immobilizing various reaction components are known in the art as described elsewhere herein.

[00133] In certain aspects, the methods for detection of nucleic acid damage can be used to test various elements of an experimental system to identify sources of such damage. For example, various buffer conditions or other components of an analytical reaction (e.g., reaction components or radiation that can induce production of oxygen radicals) can be tested to identify those that cause the least amount of damage for use in an experimental system. Further, such damage can be intentionally introduced into a nucleic acid template by the practitioner, e.g., at one or more specific locations in a template. This provides a means for controlling the progress of the polymerase, and therefore controlling the timing of production of sequence reads from

different portions of the template. For example, if the template is extremely long (e.g., thousands or tens of thousands of base pairs in length), it may be beneficial to temporarily pause the reaction at one or more points on the template to allow orientation of the sequence read to the template. In particular, a pause in emission of signal pulses is indicative that the polymerase has reached a particular location on the template, and the investigator can reinitiate polymerization by addition of repair agents/proteins to the reaction mixture. Such repair agents may be washed out of the reaction mixture and, optionally, reintroduced at a later point during the course of the reaction, e.g., by buffer exchange.

VIII. Data analysis

[00134] Analysis of the data generated by the methods described herein is generally performed using software and/or statistical algorithms that perform various data conversions, e.g., conversion of signal emissions into basecalls, conversion of basecalls into consensus sequences for a nucleic acid template, and conversion of various aspects of the basecalls and/or consensus sequence to derive a reliability metric for the resulting values. Such software, statistical algorithms, and use thereof are described in detail, e.g., in U.S. Patent Publication No. 20090024331 and U.S.S.N. 61/116,439, the disclosures of which are incorporated herein by reference in their entireties for all purposes. Specific methods for discerning altered nucleotides in a template nucleic acid are provided in U.S.S.N. 61/201,551, filed December 11, 2008, and incorporated herein by reference in its entirety for all purposes. These methods include use of statistical classification algorithms that analyze the signal from a single-molecule sequencing technology and detect significant changes in one or more aspects of signal morphology, variation of reaction conditions, and adjustment of data collection parameters to increase sensitivity to changes in signal due to the presence of modified or damaged nucleotides.

[00135] In certain aspects, the invention provides methods for detecting changes in the kinetics (e.g., slowing or pausing) or other reaction data for real-time DNA sequencing. As discussed at length above, detection of a change in such sequencing applications can be indicative of secondary structure in the template, the presence of modifications in the template, the presence of an agent bound to the template, and the like. It is appreciated that the kinetic activity of single molecules does not follow the regular and simple picture implied by traditional chemical kinetics, a view dominated by single-rate exponentials and the smooth results of ensemble averaging. In a large multi-dimensional molecular system, such as the polymerase-DNA complex, there are processes taking place on many different time scales, and the resultant kinetic picture can be quite complex at the molecular level. (See, e.g., Herbert, et al. (2008) Ann Rev Biochem 77:149.) As such, a real-time single-molecule sequencing technology should be

adaptable to such non-exponential behavior. For example, pauses during a real-time sequencing reaction are detectable as regions in the trace of observed signals over time in which it appears that the enzyme has significantly slowed as compared to the average rate of incorporation. As such, methods are provided to analyze the data generated in the vicinity of a pause site, and in particular algorithmic methods for classifying and removing or down-weighting the occurrence of pauses in the context of single-molecule sequencing. General information on algorithms for use in sequence analysis can be found, e.g., in Braun, et al. (1998) Statist Sci 13:142; and Durbin, et al. (1998) *Biological sequence analysis: Probabilistic models of proteins and nucleic acids*, Cambridge University Press: Cambridge, UK.

[00136] In certain preferred embodiments, the methods utilize a segmentation algorithm for discriminating pause regions in a real-time signal generated by monitoring single-molecule kinetics, in particular by monitoring DNA synthesis by DNA polymerase. The central observation is that during a pause the density of signal events (incorporations) is lowered, where the density refers to the number of events per a fixed unit of time. At the same time, stochastic events arising from Poisson processes, such as sticks (signals that do not correspond to an incorporation event, e.g., dyes that enter the detection volume but are not linked to nucleotides that are incorporated into the nascent strand) should continue at the same density as normally observed. Figure 7 illustrates an observation of true incorporations (solid line) versus stochastic pulses (dashed line) across time. A pause is identified in the region of the trace in which the observation of true incorporations dips below the observation of stochastic pulses. As such, by observation of differences in local densities, pauses in incorporation activity can be identified.

[00137] Other features that are related to pausing and can contribute to a full model of the phenomenon. In particular, the local sequence context of the template strand can also influence and inhibit the activity of the polymerase along the template. For example, the local sequence context that influences and/or inhibits activity of the polymerase may extend for at least about one, two, three, four, five, seven, ten, fifteen, twenty nucleotide positions, and these positions may lie upstream or downstream of the modification, or may flank the modification in the template. Other known models can also be used in the methods described herein, as will be clear to one of ordinary skill upon review of the teachings herein.

[00138] In a preferred embodiment, an algorithm for use in detecting changes during template-directed nucleic acid synthesis comprises the following general steps. First, a classifier is created that can distinguish between true incorporations and stochastic pulses. Features that can help discriminate between the two include, e.g., pulse height, pulse width, local signal-to-noise ratio, dye channel, and the χ^2 metric for the measured spectrum. Many different statistical classification algorithms known in the art can be used in this classifier. Certain preferred

algorithms include classification-and-regression trees (CART), naïve Bayesian classifiers, kernel density methods, linear discriminant functions, and neural networks. Further, the pulse classifier does not need to be particularly powerful (in an optimal specificity/sensitivity sense), because a strength of the approach relies on the greater significance associated with observing clusters of weakly significant events.

[00139] A second step is to slide a fixed-length window across the observed signal trace and count the number of incorporations versus stochastic pulses in each window using the classifier. Choice of the window size is determined by the length scale of events to be detected; a reasonable choice in practice is 10 seconds, but a practioner may increase or decrease the window size according to a particular implementation of the invention. Regions of the trace in which the stochastic pulse density exceeds the incorporation density for an extended period of time, e.g., 5-15 seconds, or more preferably about 10 seconds, are identified as corresponding to a likely pause site. These regions can be found using standard peak-finding techniques, e.g., threshold detection, finite-state machines, multi-scale methods, etc. This method has the further advantage of focusing on pause regions that have detrimental effects in the downstream use of sequencing data. The mere occurrence of a pause during sequencing is not consequential to the use of the data for DNA sequence analysis, however the occurrence of a large number of stochastic pulses in the pause region does complicate the use of the resulting data.

[00140] A variation on this exemplary algorithm is to use the time between true pulses that are identified by the classifier as a discriminator for finding pause regions, where regions that have a large difference between true pulses are candidate regions for pauses. For example, A plot of Δt (the time between true pulses) versus time will have local maxima at the location of candidate pauses.

[00141] A more sophisticated algorithm for use in detecting pause regions is a segmenting algorithm based on a hidden Markov model (HMM) architecture. Figure 8 provides an illustrative example of a simple hidden Markov model for classifying pause (P) versus sequencing (S) states within a sequencing trace. The use of this model assumes that each pulse can be labeled as either a probable incorporation (A_S, C_S, G_S, T_S) or a stochastic pulse (A_P, C_P, G_P, T_P). By fitting this model on multiple instances of sequence data (using e.g. the Baum-Welch algorithm), good emission and transition probabilities that correspond to the hidden pause and sequencing states can be generated. When subsequently presented with any particular signal (observed labels), the model can be queried for the underlying sequencing of hidden states using the Viterbi algorithm. This model is more powerful than the more simple algorithmic approaches suggested above in several ways. First, this model permits the modeling of per-nucleotide likelihoods for incorporations or sticks during the pauses or sequencing states. An

example where this is useful is if a stick/incorporation classifier for one nucleotide is particularly effective and if the pulses for another nucleotide are difficult to classify in this way. This model permits some nucleotide-specific differences in the classification power for stochastic pulses versus incorporations. A further advantage of this approach is that it is more adaptable to detecting regions across multiple time scales, where HMM segmentation approaches are usually better able to handle multi-time scale classification. The final assignment of pause regions is made by computing the log-odds ratio $\log\frac{p\langle P|x_i\rangle}{p\langle S|x_i\rangle}$ across the pulses (x_i) and identifying regions of high pause likelihood.

[00142] A more powerful algorithmic architecture for segmentation is the use of the conditional random field framework (CRF). The object is to predict the conditional probability of a signal arising from a pause or sequencing state given the observed pulses: $p\langle y|x\rangle = \frac{\exp\left[w^T F(x,y)\right]}{\sum_{y'}\exp\left[w^T F(x,y')\right]}$, where y is the sequence of desired labels (pause, sequencing), x is the observed pulse data (both basecalls and other pulse features), w is the weight vector learned from the training data, and the F function is the feature vector. The weights in the CRF can be trained using labeled sequences using standard techniques from the CRF literature (for example, Lafferty et al (2001) Proc. 18th International Conference on Machine Learning, 282-289, which is incorporated herein by reference in its entirety for all purposes). By labeling the sequence in this way, regions of high pause probability can be identified using the methods described above. One advantage of this method is the lack of a requirement for a per-pulse classifier for distinguishing between incorporation and stochastic pulses. It can also better integrate knowledge of the inter-pulse spacing and other information, such as sequence context, into a broad model. Potential disadvantages are the large amount of training data required to build the model and the algorithmic complexity involved in constructing a CRF model.

[00143] The application of such algorithmic methods to identify pause sites and/or regions with locally high stochastic pulses in sequencing trace data is useful in a number of contexts. For example, pulses in regions that are predicted to exhibit enzyme pausing can be labeled as less confident (lower quality value) for their use in downstream analyses such as sequence variant detection in resequencing applications or overlap detection for *de novo* assembly. In other embodiments, pulses from regions with a high probability of containing a cluster of stochastic pulses can be removed from the reported basecalled sequence, thereby improving the accuracy of sequence data for downstream use without resorting to secondary information such as quality values. In other embodiments, the occurrence of pauses can be

associated with other observables of interest, such as the probable DNA sequence or the occurrence of modified nucleotide bases. For example, sequences upstream of a pause site can be called in part based on their known effect on pausing. That is, if a pause occurs downstream of a sequence, then the sequence is more likely to be one that facilitates or exacerbates pausing than one that has no effect or that reduces the likelihood of pausing. As such, if sequencing of a modification is known to increase the likelihood of pausing, then this information can be incorporated into a Bayesian likelihood model for identifying modified bases. In further embodiments, the pause detection methods described herein can also be used to increase the understanding of the biophysics of polymerase activity, thereby providing useful feedback to efforts to better develop single-molecule, real-time sequencing techniques.

[00144] Algorithms for the identification of regions in sequence data belong to the general category of sequence labeling or segmentation algorithms, which are generally known in the art. The mapping of this problem to sliding-window analysis, HMMs, or CRFs is natural in this context. Other algorithms that approach the same problem are multiple change-point analysis such as the Gibbs sampler (see, e.g., Lee, P.M. (2004) *Bayesian Statistics: An Introduction*, Oxford University Press: New York, NY, the disclosure of which is incorporated herein by reference in its entirety for all purposes), or locally weighted polynomial regression (see, e.g., Braun, et al., *supra*).

[00145] In general, data analysis methods benefit when the sequencing technology generates redundant sequence data for a given template molecule, e.g. by molecular redundant sequencing as described above. The distribution of IPDs for each read at that position is an exponential. The decay constant for the exponential of a methylated base and for that of an unmethylated base may be different. However, because of the large amount of overlap between two exponentials, it is still challenging to use one read to distinguish between the two populations. However, if one takes the mean of multiple reads at a single position, the distribution of this mean is a gamma function (convolution of several exponentials), which is more Gaussian-like and better separated than exponentials. This enables better distinguishability of the two populations. For example, Figure 13 provides actual data showing that for two different positions in a single circular template, one always unmethylated, and one differentially methylated, an increase in the number of reads for the template corresponds to an increased resolution between IPDs for methylated vs. unmethylated adenosines. If the underlying distributions are exponential, as just discussed, then the mean value is the only metric that can be used for making the distinction (the standard deviation is the same as the mean). If the distribution is non-exponential for each read position, as it would be for the methylcytosine IPD' that is weighted over numerous neighboring positions and thus itself has a gamma-like

distribution, then when doing consensus reads of the same position, one can take into account the mean of the gamma-like weighted IPD' distributions along with other information, e.g. its standard deviation, its skewness, or other characteristics of the distribution. Figure 10 shows actual molecular consensus distributions for methylcytosine, given the underlying gamma-like weighted IPD' distributions of individual reads, but in this figure only the means of these underlying distributions were utilized. The plotted distributions could become even more well-separated if other characteristics had been taken into account. The data used to generate Figures 13 and 10 is more fully described in the Examples herein.

[00146] In certain embodiments, methods may be employed that use weighted sums of signal features at multiple positions to determine the status of a base, e.g., whether or not it is methylated in a template nucleic acid. In particular, interpulse duration (IPD) information from multiple positions can be used to determine whether or not a given cytosine is methylated, e.g., by comparing nascent strand synthesis data for a differentially methylated template (Me+) to such data for a fully unmethylated template (Me-). In certain preferred embodiments, a pseudo IPD is created for a given template position that is actually a weighted sum of the IPDs for the surrounding positions. More specifically, $IPD'_j = \sum_i w_i \times \frac{IPD_{ji}}{\langle IPD_j \rangle_{Me-}}$, where j is the index of the cytosine in question; i is an index that ranges over all the neighboring positions that yield a change in IPD due to cytosine being either methylated or unmethylated; and $\langle IPD_i \rangle_{Me-}$ is the average IPD at that particular position in the Me- template. The individual weights for the multiple positions (all together which would likely sum to 1) could be based on a combination of the following metrics, assuming we are comparing two templates that are identical aside for one being methylated and the other being unmethylated at a given position: the ratio of or difference in IPD between the two templates at that given position; the statistical significance of the distinguishability between the IPD distributions of the two templates at that given positions; the number of observations used when creating the IPD distributions; and the neighboring sequence context. An example w_i could be $w_i = \log\left[\frac{\langle IPD_i \rangle_{Me+}}{\langle IPD_i \rangle_{Me-}}\right]$. This signal can also be weighted by the prior probability of seeing a Me+ signal.

[00147] In certain aspects, the invention provides a general-purpose approach to discriminating between Me+ and Me- using features in a real-time sequencing-by-synthesis trace comprising signals emitted during the incorporation of optically detectable nucleotides into a nascent strand by a polymerase enzyme. Such traces and various methods of analysis thereof are further described elsewhere, e.g., in U.S. Patent Publication No. 20090024331, incorporated

herein by reference in its entirety for all purposes. A first stage of this approach includes the development of a classifier for distinguishing methylated from unmethylated cytosines in a nucleic acid template. A set of features is measured for every pulse (discrete event or signal) in the trace. For example, a set of measurable features might be {pulse width, pulse duration, pulse height, pulse amplitude variability}. Call the values of these features $f_1^i, f_2^i, \ldots, f_k^i$ for pulse *i*. For each cytosine pulse in the methylated and unmethylated template data sets (which may or may not be restricted to CpG), tabulation is performed for the local pulse features $\vec{f} = \{f_1^{i-3}, f_2^{i-3}, \ldots, f_k^{i-1}, f_1^i, f_2^i, \ldots, f_k^i, \ldots, f_1^{i+3}, f_2^{i+3}, \ldots, f_k^{i+3}\}$.

In this example, a local context extending 3 pulses to the left and right of the pulse of interest is assumed, but this context size is flexible and, in certain embodiments, can be application-specific.

[00148] The observed data likelihoods $p(\vec{f}|Me+)$ and $p(\vec{f}|Me-)$ are derived, e.g., by a kernel density method or simple binning and tabulation of the features. Thus, a generalized signal for determining methylation status on the trace has been determined:

$w_i = \log\left[\frac{p(\vec{f}|Me+)p(Me+)}{p(\vec{f}|Me-)p(Me-)}\right]$, where $p(Me+)$ and $p(Me-)$ are the prior probabilities of methylated or un-methylated positions, respectively.

[00149] Various standard classification algorithm development techniques known to those of ordinary skill in the art may be applied to refine this approach, both to reduce training set bias and to improve sensitivity. Such techniques include but are not limited to cross-validation, boosting, and bootstrap aggregating (bagging). In certain embodiments, the set of feature inputs is restricted to those that are most correlated with the Me+ and Me- status of a position. In certain embodiments, the major component in a principal components analysis can serve as a better weighted combination of the most important features. In further embodiments, leave-one-out cross-validation can be valuable in selecting a robust predictive algorithm, e.g., by mitigating overfitting to the observed data that can occur when developing a classifier on a training set. Further, in some embodiments a boosting approach (training of a hierarchy of classifiers on the progressively more difficult regions of feature space) is applied to improve sensitivity.

[00150] More sophisticated signals can be employed to detect multiple, closely spaced CpGs. In certain embodiments, the data likelihoods described above can be measured for the case of two CpG sites with a known methylation state located a known distance apart, e.g., 2 base pairs apart. The signal generalizes to

$$w_i = \underset{\mu\in\{++,+-,-+,--\}}{\arg\max} \log\left[\frac{p(\vec{f}_\alpha,\vec{f}_\beta|\mu)p(\mu)}{\sum_{\mu'} p(\vec{f}_\alpha,\vec{f}_\beta|\mu')p(\mu')}\right],$$

where $p(\mu)$ and $p(\mu')$ are the prior probabilities of (methylated,un-methylated) configurations (this joint distribution would be assumed to be independent unless otherwise shown).

[00151] Although described primarily in the context of detection of methyl cytosine, these methods are also applicable to methyl-adenosine or any other base modification for which IPDs are used as a metric for detection. Figure 14 provides data showing differences between ratios of IPDs for methylated adenosines and unmethylated adenosines in a template nucleic acid, and the data used to generate Figure 14 is further described in the Examples herein. This data also shows that N^6-methyladenosine, like methylcytosine, has an effect on IPD not only at the methylated base but also at multiple, neighboring positions, as well. Further, in light of the above teachings it will be clear to one of ordinary skill that the approach can be extended to pulse metrics other than IPD, such as pulse width, branch rate, mismatch rate, deletion rate, etc. In addition, the general classifier approach suggested in steps 2+3 can be implemented with many standard statistical classification algorithms, i.e. linear discriminant analysis, multi-dimensional regression, kernel methods, classification and regression trees, neural networks, and support vector machines. The approach can also incorporate data from multiple strands of a duplex template. For example, because the CG sequence for cytosine methylation and the GATC sequence for adenosine methylation is the same on the reverse complement strand, these bases can be methylated on both complementary strands. If the general statistical distribution for the fraction of sites that are hemi-methylated vs. fully methylated is known, then information regarding IPD or other metrics gained from the complementary strand can be used to increase the accuracy with which a call is made on a particular strand. For example, if after analyzing each strand separately it is concluded that there is a 95% chance that stand A is methylated and a 55% chance that complementary strand B is methylated, but it is known that there is a 80% chance that if one strand is methylated then so is the other, then the confidence in calling strand B as methylated is increased.

[00152] Another modified base for which IPDs may be used as a metric for detection is 5-hydroxymethylcytosine (5-hmC). It was recently found to be abundant in human and mouse brains, as well as in embryonic stem cells (see, e.g., Kriaucionis, et al. (2009) "The nuclear DNA base 5-hydroxymethylcytosine is present in Purkinje neurons and the brain" *Science* **324** (5929): 929–30; and Tahiliani M et al. (May 2009) "Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1" *Science* **324** (5929): 930–

35, incorporated herein by reference in their entireties for all purposes). In mammals, it can be generated by oxidation of 5-methylcytosine, a reaction mediated by the Tet family of enzymes. Conventional bisulfite sequencing does not effectively distinguish 5-hmC from 5-MeC because 5-hmC tends to remain unmodified like 5-MeC. As such, mass spectrometry is the typical means of detecting 5-hmC in a nucleic acid sample. The methods described herein provide a high-throughput, real-time method to distinguish between C, 5-MeC, and 5-hmC by monitoring deviations from normal polymerase kinetics, including IPD and pulse width.

[00153] Experiments were carried out to test the ability of the methods of the invention to distinguish between 5-MeC and 5-hmC, and it was found that 5-hmC causes an increase in IPD at certain positions surrounding the 5-hmC site in the template, and also decreases the pulse width at that position. Further, the data generated suggests that 5-hmC may also increase the pulse width at the position following the 5-hmC site. Both the difference in IPD and the difference in pulse width between C and 5-hmC were larger in magnitude than were the differences in IPD and pulse width between C and 5-MeC, and these larger magnitudes are likely to make 5-hmC even more detectable than 5-MeC. Without being bound by theory, the reason for the higher magnitude differences for these two measures may be due to the additional oxygen atom present in 5-hmC as compared to 5-MeC. This additional oxygen could yield additional steric and charge-based interactions between the polymerase and the DNA template that slow the binding and/or incorporation of the complementary base into the nascent strand.

[00154] Based on the findings that indicate the easier detection of 5-hmC as compared to 5-MeC, in certain embodiments template nucleic acids can be treated to convert 5-MeC to the more easily detected 5-hmC, e.g., by treatment with an enzyme such as TET1, which converts 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA (see, e.g., Tahiliani M et al., *supra*). Although this technique would not permit distinction between 5-MeC and 5-hmC in the template (since the 5-MeC converted to 5-hmC would be indistinguishable from any 5-hmC originally present in the template), it will nonetheless be useful for facilitating detection 5-MeC patterns in template nucleic acids with the caveat that the patterns so discovered may, in vivo, also include 5-hmC bases.

[00155] In order to maximally use the IPD and pulse width signals from multiple positions surrounded the 5-hmC site, one could use a technique to find the optimal weighting of different positions for IPD and pulse width in order to distinguish 5-hmC, 5-MeC, and C from one another. An example of one such technique is principle component analysis, and others are known in the art. Principle component analysis can be described as finding the eigenvector (using each metric such as IPD or pulse width at each position as a different basis vector, such that if you have 10 positions in question and two metrics, your basis will have 2x10=20

dimensions) with the greatest eigenvalue. For a review of principle component analysis, see e.g. Jolliffe I.T. Principal Component Analysis, Series: Springer Series in Statistics, 2nd ed., Springer, NY, 2002, the disclosure of which is incorporated herein by reference in its entirety for all purposes.

IX. Systems

[00156] The invention also provides systems that are used in conjunction with the compositions and methods of the invention in order to provide for real-time single-molecule detection of analytical reactions. In particular, such systems typically include the reagent systems described herein, in conjunction with an analytical system, e.g., for detecting data from those reagent systems. In certain preferred embodiments, analytical reactions are monitored using an optical system capable of detecting and/or monitoring interactions between reactants at the single-molecule level. For example, such an optical system can achieve these functions by first generating and transmitting an incident wavelength to the reactants, followed by collecting and analyzing the optical signals from the reactants. Such systems typically employ an optical train that directs signals from the reactions to a detector, and in certain embodiments in which a plurality of reactions is disposed on a solid surface, such systems typically direct signals from the solid surface (e.g., array of confinements) onto different locations of an array-based detector to simultaneously detect multiple different optical signals from each of multiple different reactions. In particular, the optical trains typically include optical gratings or wedge prisms to simultaneously direct and separate signals having differing spectral characteristics from each confinement in an array to different locations on an array based detector, e.g., a CCD, and may also comprise additional optical transmission elements and optical reflection elements.

[00157] An optical system applicable for use with the present invention preferably comprises at least an excitation source and a photon detector. The excitation source generates and transmits incident light used to optically excite the reactants in the reaction. Depending on the intended application, the source of the incident light can be a laser, laser diode, a light-emitting diode (LED), a ultra-violet light bulb, and/or a white light source. Further, the excitation light may be evanescent light, e.g., as in total internal reflection microscopy, certain types of waveguides that carry light to a reaction site (see, e.g., U.S. Application Pub. Nos. 20080128627, 20080152281, and 200801552280), or zero mode waveguides, described below. Where desired, more than one source can be employed simultaneously. The use of multiple sources is particularly desirable in applications that employ multiple different reagent compounds having differing excitation spectra, consequently allowing detection of more than one fluorescent signal to track the interactions of more than one or one type of molecules simultaneously (e.g., multiple types of

differentially labeled reaction components). A wide variety of photon detectors or detector arrays are available in the art. Representative detectors include but are not limited to an optical reader, a high-efficiency photon detection system, a photodiode (e.g. avalanche photo diodes (APD)), a camera, a charge-coupled device (CCD), an electron-multiplying charge-coupled device (EMCCD), an intensified charge coupled device (ICCD), and a confocal microscope equipped with any of the foregoing detectors. For example, in some embodiments an optical train includes a fluorescence microscope capable of resolving fluorescent signals from individual sequencing complexes. Where desired, the subject arrays of optical confinements contain various alignment aides or keys to facilitate a proper spatial placement of the optical confinement and the excitation sources, the photon detectors, or the optical train as described below.

[00158] The subject optical system may also include an optical train whose function can be manifold and may comprise one or more optical transmission or reflection elements. Such optical trains preferably encompass a variety of optical devices that channel light from one location to another in either an altered or unaltered state. First, the optical train collects and/or directs the incident wavelength to the reaction site (e.g., optical confinement). Second, it transmits and/or directs the optical signals emitted from the reactants to the photon detector. Third, it may select and/or modify the optical properties of the incident wavelengths or the emitted wavelengths from the reactants. Illustrative examples of such optical transmission or reflection elements are diffraction gratings, arrayed waveguide gratings (AWG), optical fibers, optical switches, mirrors (including dichroic mirrors), lenses (including microlenses, nanolenses, objective lenses, imaging lenses, and the like), collimators, optical attenuators, filters (e.g., polarization or dichroic filters), prisms, wavelength filters (low-pass, band-pass, or high-pass), planar waveguides, wave-plates, delay lines, and any other devices that guide the transmission of light through proper refractive indices and geometries. One example of a particularly preferred optical train is described in U.S. Patent Pub. No. 20070036511, filed August 11, 2005, and incorporated by reference herein in its entirety for all purposes.

[00159] In a preferred embodiment, a reaction site (e.g., optical confinement) containing a reaction of interest is operatively coupled to a photon detector. The reaction site and the respective detector can be spatially aligned (e.g., 1:1 mapping) to permit an efficient collection of optical signals from the reactants. In certain preferred embodiments, a reaction substrate is disposed upon a translation stage, which is typically coupled to appropriate robotics to provide lateral translation of the substrate in two dimensions over a fixed optical train. Alternative embodiments could couple the translation system to the optical train to move that aspect of the system relative to the substrate. For example, a translation stage provides a means of removing a

reaction substrate (or a portion thereof) out of the path of illumination to create a non-illuminated period for the reaction substrate (or a portion thereof), and returning the substrate at a later time to initiate a subsequent illuminated period. An exemplary embodiment is provided in U.S. Patent Pub. No. 20070161017, filed December 1, 2006.

[00160] In particularly preferred aspects, such systems include arrays of reaction regions, e.g., zero mode waveguide arrays, that are illuminated by the system, in order to detect signals (e.g., fluorescent signals) therefrom, that are in conjunction with analytical reactions being carried out within each reaction region. Each individual reaction region can be operatively coupled to a respective microlens or a nanolens, preferably spatially aligned to optimize the signal collection efficiency. Alternatively, a combination of an objective lens, a spectral filter set or prism for resolving signals of different wavelengths, and an imaging lens can be used in an optical train, to direct optical signals from each confinement to an array detector, e.g., a CCD, and concurrently separate signals from each different confinement into multiple constituent signal elements, e.g., different wavelength spectra, that correspond to different reaction events occurring within each confinement. In preferred embodiments, the setup further comprises means to control illumination of each confinement, and such means may be a feature of the optical system or may be found elsewhere is the system, e.g., as a mask positioned over an array of confinements. Detailed descriptions of such optical systems are provided, e.g., in U.S. Patent Pub. No. 20060063264, filed September 16, 2005, which is incorporated herein by reference in its entirety for all purposes.

[00161] The systems of the invention also typically include information processors or computers operably coupled to the detection portions of the systems, in order to store the signal data obtained from the detector(s) on a computer readable medium, e.g., hard disk, CD, DVD or other optical medium, flash memory device, or the like. For purposes of this aspect of the invention, such operable connection provides for the electronic transfer of data from the detection system to the processor for subsequent analysis and conversion. Operable connections may be accomplished through any of a variety of well known computer networking or connecting methods, e.g., Firewire®, USB connections, wireless connections, WAN or LAN connections, or other connections that preferably include high data transfer rates. The computers also typically include software that analyzes the raw signal data, identifies signal pulses that are likely associated with incorporation events, and identifies bases incorporated during the sequencing reaction, in order to convert or transform the raw signal data into user interpretable sequence data (see, e.g., Published U.S. Patent Application No. 2009-0024331, the full disclosure of which is incorporated herein by reference in its entirety for all purposes).

[00162] Exemplary systems are described in detail in, e.g., U.S. Patent Application No. 11/901,273, filed September 14, 2007 and U.S. Patent Application No. 12/134,186, filed June 5, 2008, the full disclosures of which are incorporated herein by reference in their entirety for all purposes.

[00163] Further, the invention provides data processing systems for transforming raw data generated in an analytical reaction into analytical data that provides a measure of one or more aspects of the reaction under investigation, e.g., transforming signals from a sequencing-by-synthesis reaction into nucleic acid sequence read data, which can then be transformed into consensus sequence data. In certain embodiments, the data processing systems include machines for generating nucleic acid sequence read data by polymerase-mediated processing of a template nucleic acid molecule (e.g., DNA or RNA). The nucleic acid sequence read data generated is representative of the nucleic acid sequence of the nascent polynucleotide synthesized by a polymerase translocating along a nucleic acid template only to the extent that a given sequencing technology is able to generate such data, and so may not be identical to the actual sequence of the nascent polynucleotide molecule. For example, it may contain a deletion or a different nucleotide at a given position as compared to the actual sequence of the polynucleotide, e.g., when a nucleotide incorporation is missed or incorrectly determined, respectively. As such, it is beneficial to generate redundant nucleic acid sequence read data, and to transform the redundant nucleic acid sequence read data into consensus nucleic acid sequence data that is generally more representative of the actual sequence of the polynucleotide molecule than nucleic acid sequence read data from a single read of the nucleic acid molecule. Redundant nucleic acid sequence read data comprises multiple reads, each of which includes at least a portion of nucleic acid sequence read that overlaps with at least a portion of at least one other of the multiple nucleic acid sequence reads. As such, the multiple reads need not all overlap with one another, and a first subset may overlap for a different portion of the nucleic acid sequence than does a second subset. Such redundant sequence read data can be generated by various methods, including repeated synthesis of nascent polynucleotides from a single nucleic acid template, synthesis of polynucleotides from multiple identical nucleic acid templates, or a combination thereof.

[00164] In another aspect, the data processing systems can include software and algorithm implementations provided herein, e.g. those configured to transform redundant nucleic acid sequence read data into consensus nucleic acid sequence data, which, as noted above, is generally more representative of the actual sequence of the nascent polynucleotide molecule than nucleic acid sequence read data from a single read of a single nucleic acid molecule. Further, the transformation of the redundant nucleic acid sequence read data into consensus nucleic acid sequence data identifies and negates some or all of the single-read variation

between the multiple reads in the redundant nucleic acid sequence read data. As such, the transformation provides a representation of the actual nucleic acid sequence of the nascent polynucleotide complementary to the nucleic acid template that is more accurate than a representation based on a single read.

[00165] Various methods and algorithms for data transformation employ data analysis techniques that are familiar in a number of technical fields, and are generally referred to herein as statistical analysis. For clarity of description, details of known techniques are not provided herein. These techniques are discussed in a number of available reference works, such as those provided in U.S. Patent Publication No. 20090024331 and U.S.S.N. 61/116,439, filed November 20, 2008, the disclosures of which are incorporated herein by reference in their entireties for all purposes.

[00166] The software and algorithm implementations provided herein are preferably machine-implemented methods, e.g., carried out on a machine comprising computer-readable medium configured to carry out various aspects of the methods herein. For example, the computer-readable medium preferably comprises at least one or more of the following: a) a user interface; b) memory for storing raw analytical reaction data; c) memory storing software-implemented instructions for carrying out the algorithms for transforming the raw analytical reaction data into transformed data that characterizes one or more aspects of the reaction (e.g., rate, consensus sequence data, etc.); d) a processor for executing the instructions; e) software for recording the results of the transformation into memory; and f) memory for recordation and storage of the transformed data. In preferred embodiments, the user interface is used by the practitioner to manage various aspects of the machine, e.g., to direct the machine to carry out the various steps in the transformation of raw data into transformed data, recordation of the results of the transformation, and management of the transformed data stored in memory.

[00167] As such, in preferred embodiments, the methods further comprise a transformation of the computer-readable medium by recordation of the raw analytical reaction data and/or the transformed data generated by the methods. Further, the computer-readable medium may comprise software for providing a graphical representation of the raw analytical reaction data and/or the transformed data, and the graphical representation may be provided, e.g., in soft-copy (e.g., on an electronic display) and/or hard-copy (e.g., on a print-out) form.

[00168] The invention also provides a computer program product comprising a computer-readable medium having a computer-readable program code embodied therein, the computer readable program code adapted to implement one or more of the methods described herein, and optionally also providing storage for the results of the methods of the invention. In certain

preferred embodiments, the computer program product comprises the computer-readable medium described above.

[00169] In another aspect, the invention provides data processing systems for transforming raw analytical reaction data from one or more analytical reactions into transformed data representative of a particular characteristic of an analytical reaction, e.g., an actual sequence of one or more template nucleic acids analyzed, a rate of an enzyme-mediated reaction, an identity of a kinase target molecule, and the like. Such data processing systems typically comprise a computer processor for processing the raw data according to the steps and methods described herein, and computer usable medium for storage of the raw data and/or the results of one or more steps of the transformation, such as the computer-readable medium described above.

[00170] As shown in Figure 9, the system 900 includes a substrate 902 that includes a plurality of discrete sources of chromophore emission signals, e.g., an array of zero mode waveguides 904. An excitation illumination source, e.g., laser 906, is provided in the system and is positioned to direct excitation radiation at the various signal sources. This is typically done by directing excitation radiation at or through appropriate optical components, e.g., dichroic 908 and objective lens 910, that direct the excitation radiation at the substrate 902, and particularly the signal sources 904. Emitted signals from the sources 904 are then collected by the optical components, e.g., objective 910, and passed through additional optical elements, e.g., dichroic 908, prism 912 and lens 914, until they are directed to and impinge upon an optical detection system, e.g., detector array 916. The signals are then detected by detector array 916, and the data from that detection is transmitted to an appropriate data processing system, e.g., computer 918, where the data is subjected to interpretation, analysis, and ultimately presented in a user ready format, e.g., on display 920, or printout 922, from printer 924. As will be appreciated, a variety of modifications may be made to such systems, including, for example, the use of multiplexing components to direct multiple discrete beams at different locations on the substrate, the use of spatial filter components, such as confocal masks, to filter out-of focus components, beam shaping elements to modify the spot configuration incident upon the substrates, and the like (See, e.g., Published U.S. Patent Application Nos. 2007/0036511 and 2007/095119, and U.S. Patent Application No. 11/901,273, all of which are incorporated herein by reference in their entireties for all purposes.)

II. Examples

Detection of 5-methylcytosine (5-MeC)

[00171] Methylation sequencing on a SMRT™ Sequencing platform (see, e.g., P. M. Lundquist, et al., *supra*) was performed on short, synthetic DNA oligos with contrived patterns

of methylated and unmethlyated bases, along with control sequences having the same primary sequence but without any methylation. These templates provided unequivocal fluorescence pulse patterns and tempos that demonstrated how the combination of sequence context and methylation status affected interpulse duration. For example, SMRT™ sequencing experiments were performed using synthetic DNA templates that only differed by a single methylated *vs.* unmethylated cytosine. The difference in average interpulse durations between the two templates was visible both at the 5-MeC position and in the vicinity of the 5-MeC position.

[00172] Because the interpulse duration between any two successive incorporation events is stochastic in nature and has an exponential distribution (Eid, et al., *supra*), a single sequencing measurement may not always yield enough information to determine methylation status with certainty. Therefore, in certain embodiments a highly processive, strand-displacing polymerase is used, and this polymerase carries out multiple laps of synthesis around a circular DNA template (J. Korlach, et al., *Proc Natl Acad Sci U S A* **2008**, *supra*). This mode of operation provides repeated sequencing of the same DNA molecule to generate multiple sequence reads, e.g., by rolling circle replication. The statistical distribution of interpulse durations obtained at a particular template site will thus indicate its methylation state.

[00173] In particular, Figure 10A shows a schematic of two templates for use in SMRT™ sequencing. Both are comprises a double-stranded region flanked by two single-stranded hairpins. A polymerase binds to a primed location on the template, e.g., via a primer hybridized to one of the single-stranded hairpins, and commences processing the template to generate a nascent strand complementary to the strand upon which the polymerase is translocating. The strand displacement activity of the polymerase permits passage through the double-stranded region which is unwound to transform the template into a circular form. The polymerase then proceeds around the other single-stranded hairpin and on through the previously displaced strand of the double-stranded region. The polymerase can continue to process the template in a "rolling-circle" fashion to generate a concatemer comprising multiple copies of complements to both strands of the double-stranded region, as well as the hairpins. The two templates are identical except at position 2, where the top template comprises a methylated cytosine (5-MeC) and the bottom template comprises a non-methylated cytosine. (Position 1 is a non-methylated cytosine in both templates.) Figure 10B provides an illustrative depiction of the difference in IPD for the methylated template as compared to the unmethylated template. For each row, the histograms depict the distributions of mean weighted IPD (averaged over the labeled number of circular consensus sequencing subreads (in this context, a sequence read generated from a single pass of the polymerase around the template). Specifically, "1" indicates the sequencing data was derived from a sequence read generated in a single pass around the template; "3" indicates the

data was derived from a sequence read generated in three passes around the template; and "5" indicates the data was derived from a sequence read generated in five passes around the template. The data from the methylated template is shown as a solid line, and the data from the unmethylated template is shown as a dotted line. At Position 1, the distributions of weighted IPD for the two templates are very similar. At Position 2, the average weighted IPD after a single subread (top histogram) is longer in the methylated template than in the unmethylated template. After 3 and 5 circular subreads, the distributions overlap even less. The interpulse duration (IPD) was clearly lengthened by the presence of 5-MeC. These results demonstrated the ability to use SMRT™ sequencing technology to perform methylation sequencing of DNA. Weighted IPDs are described elsewhere herein.

[00174] Further, methylcytosine was shown to have an effect on interpulse duration (IPD) not only at the methylated base, but over a range of several bases upstream and downstream of the position of the methylcytosine. Specifically, an increase in IPD was observed at some positions in the presence of methylcytosine relative to the same position in the absence of methylcytosine. Figure 11 provides a plot depicting the ratio of the average IPD in the methylated template to the average IPD in the unmethylated template, plotted versus DNA template position. The two templates are identical except for the methylated bases in the methylated template, which are indicated by arrowheads in Figure 11.

[00175] Figure 12 provides another data set illustrating the ratio of IPD for a different methylated template vs. an identical but unmethylated template as a function of position. Seven cytosines (shown with crosshatching) were differentially methylated (5-MeC) between the two templates. That data clearly showed that IPD was increased in the region comprising the methylated bases. Interestingly, the effect on IPD occurred mostly downstream of the methylated positions. As such, data from nascent strand synthesis at positions in the template that are near the differentially methylated site, in addition to the differentially methylated site itself, is useful for methylation detection during real-time nascent strand synthesis.

Detection of N6-methyladenosine (N6-MeA)

[00176] Similar methods as those used to detect 5-MeC were used to detect N6-MeA in similarly constructed template nucleic acids. Figure 13A shows a schematic of two templates, both of which comprise a double-stranded region flanked by two single-stranded hairpins. The methylated template has an A within a GATC context at Position 1 and a ^{m}A within a GATC context at Position 2, whereas the unmethylated template has an A at both positions. Otherwise, the two templates are identical. As described above, a polymerase binds to a primed location on the template and commences processing the template to generate a nascent strand, using its strand displacement activity to unwind the double-stranded region and proceed around the

template. Figure 13B shows plots of mean IPD generated from sequencing data using these two templates for varying numbers of consensus reads, as described above. The data from the methylated template is shown as a solid line, and the data from the unmethylated template is shown as a dotted line. For each row in Figure 13B, the histograms depict the distributions of mean IPD (averaged over the labeled number of consensus sequencing subreads, i.e. the number of times the polymerase made one complete pass around the template to generate a complementary nascent strand). At Position 1, the distributions of IPD for the two templates are very similar. At Position 2, the average IPD after a single subread (top histogram) is ~5× longer in the methylated template than in the unmethylated template. After 3 and 5 circular subreads, the distributions overlap even less. The interpulse duration (IPD) was clearly lengthened by the presence of N6-MeA, demonstrating that the SMRT™ sequencing technology can be used to perform methylation sequencing of DNA comprising a methylated base other than 5-MeC.

[00177] Receiver operating characteristic (ROC) curves, parameterized by IPD threshold, for assigning a methylation status to an adenosine nucleotide are provided in Figure 13C. True positive means that an ^{m}A is correctly called as ^{m}A, whereas a false positive means that an A is mistakenly called as ^{m}A. These ROC curves, based on the IPD distributions from Position 2 in Figure 13B, are shown for a single read (solid line), and for 3 (long-dashed line) or 5 (short-dashed line) molecular redundant sequencing reads produced by the polymerase processing the template one, three, or five times, respectively. The dotted horizonal line bisecting the graph depicts the ROC curve for randomly guessing the methylation status. The normalized area under the ROC curve is 0.80 after the first circular subread but increases to 0.92 and 0.96 after three and five circular subreads, respectively. In fact, after five subreads, >85% of ^{m}A bases can be detected at this template position with a false positive rate of only ~5%.

[00178] Like methylcytosine, methyladenosine was also shown to have an effect on IPD over a range of several bases upstream and downstream of the position of the methyladenosine. Specifically, an increase in IPD was observed at some positions in the presence of methyladenosine relative to the same position in the absence of methyladenosine. Figure 14 provides a plot depicting the ratio of the average IPD in the methylated template to the average IPD in the unmethylated template, plotted versus DNA template position. The two templates are identical except for the methylated bases in the methylated template, which are indicated by arrowheads in Figure 14.

Detection of 5-hydroxymethylcytosine

[00179] Similar to 5-MeC and N6-MeA, 5-hydroxymethylcytosine was also tested and shown to have an effect on IPD over a range of several bases upstream and downstream of the position of the 5-hydroxymethylcytosine. Specifically, an increase in IPD was observed at some

positions in the presence of 5-hydroxymethylcytosine relative to the same position in the absence of 5-hydroxymethylcytosine. Figure 15 provides a plot depicting the ratio of the average IPD in the hydroxymethylated template to the average IPD in the unmethylated template, plotted versus DNA template position. The two templates are identical except for the hydroxymethylcytosine in the hydroxymethylated template, which are indicated by arrowheads in Figure 15. Templates comprising 5-hydroxymethylcytosine bases were also tested and the presence of these modifications was shown to have an effect on pulse width. Figure 16 provides a plot of pulse width ratio (pulse width for methylated template divided by pulse width for unmethylated template) vs. template position where the modified positions comprise 5-hydroxymethylcytosine bases. Variably hydroxymethylated positions are indicated by the arrowheads.

Detection of 8-oxoguanosine (8-oxoG)

[00180] 8-oxoguanosine was also subjected to single molecule real-time sequencing and was shown to affect IPD both at the site of the modifications as well as at proximal unmodified positions in the template. An increase in IPD was observed at some positions in the presence of 8-oxoguanosine relative to the same position in the absence of 8-oxoguanosine. Figure 17 provides a plot depicting the ratio of the average IPD in the 8-oxoguanosine template to the average IPD in the unmodified template, plotted versus DNA template position. The two templates are identical except for the 8-oxoguanosine in the modified template, which are indicated by arrowheads in Figure 17. These data showed that, compared to G, 8-oxoG altered IPD significantly over a window of ~10 neighboring bases surrounding the 8-oxoG position. Some positions saw an increase in IPD by a factor of as much as 6.5x. Templates comprising 8-oxoG bases were also tested and the presence of these modifications was shown to have an effect on pulse width. Figure 18 provides a plot of pulse width ratio (pulse width for an 8-oxoG template divided by pulse width for template with no 8-oxoG) vs. template position where the modified positions comprise 8-oxoG bases. Variable positions are indicated by the arrowheads. Further, 8-oxoG altered pulse width over a window of 7-8 neighboring bases by as much as 40%, and such alteration included both increases and decreases in pulse width.

[00181] It is to be understood that the above description is intended to be illustrative and not restrictive. It readily should be apparent to one skilled in the art that various embodiments and-modifications may be made to the invention disclosed in this application without departing from the scope and spirit of the invention. The scope of the invention should, therefore, be determined not with reference to the above description, but should instead be determined with reference to the appended claims, along with the full scope of equivalents to which such claims are entitled. All publications mentioned herein are cited for the purpose of describing and

disclosing reagents, methodologies and concepts that may be used in connection with the present invention. Nothing herein is to be construed as an admission that these references are prior art in relation to the inventions described herein. Throughout the disclosure various patents, patent applications, and publications are referenced. To the extent not already expressly incorporated herein, all published references and patent documents referred to in this disclosure are incorporated herein by reference in their entirety for all purposes.

What is claimed is:

1. A method for identifying a modification, the method comprising:
   a) providing a template nucleic acid comprising the modification;
   b) providing an enzyme capable of processing the template nucleic acid;
   c) contacting the template nucleic acid with the enzyme;
   d) monitoring processing of the template nucleic acid by the enzyme; and
   e) detecting a change in the processing, wherein the change is indicative of the modification, thereby indentifying the modification.

2. The method of claim 1, wherein the modification is selected from the group consisting of a methylated base, a pseudouridine base, an 7,8-dihydro-8-oxoguanine base, a 2'-O-methyl derivative base, a nick, an apurinic site, an apyrimidic site, a pyrimidine dimer, a cis-platen crosslinking, oxidation damage, hydrolysis damage, a bulky base adduct, a thymine dimer, a photochemistry reaction product, an interstrand crosslinking product, a mismatched base, a secondary structure, or a bound agent.

3. The method of claim 2, wherein the modification is a methylated cytosine or adenine base.

4. The method of claim 1, wherein the template nucleic acid comprises a single-stranded portion and a double-stranded portion.

5. The method of claim 4, wherein the double-stranded portion is a result of complementarity between two separate portions of the template nucleic acid.

6. The method of claim 1, wherein the template nucleic acid comprises a first polynucleotide region comprising the modification and a second polynucleotide region complementary to the first polynucleotide region, where the first polynucleotide region and the second polynucleotide region are on a single strand of the template nucleic acid.

7. The method of claim 1, wherein the template nucleic acid is subjected to a treatment to alter the modification prior to the contacting,

8. The method of claim 7, wherein the treatment is selected from the group consisting of: glycosylase modification, bisulfite modification, DMS modification, cytosine methyltransferase modification, TET1 modification, restriction enzyme modification, and cytidine deaminase modification.

9. The method of claim 8, wherein the treatment is glycosylase modification to convert the modification to an abasic site, and further wherein the processing results in the synthesis of a

nascent nucleic acid strand into which a non-natural nucleotide analog is incorporated opposite the abasic site.

10. The method of claim 9, wherein the non-natural nucleotide analog comprises a detectable label that distinguishes its incorporation from incorporation of other nucleotides into the nascent nucleic acid strand.

11. The method of claim 9, wherein the non-natural nucleotide analog is a pyrene analog.

12. The method of claim 9, wherein the non-natural nucleotide analog is a mixture of incorporatable non-natural nucleotide analogs that are non-incorporatable non-natural nucleotide analogs.

13. The method of claim 12, wherein the non-incorporatable non-natural nucleotide analogs associates with the abasic site multiple times prior to incorporation of one of the incorporatable non-natural nucleotide analogs.

14. The method of claim 8, wherein the treatment is DMS modification and the modification is secondary structure of the template nucleic acid or an agent bound to the template nucleic acid.

15. The method of claim 1, wherein the template nucleic acid is a circular nucleic acid.

16. The method of claim 15, wherein the processing is rolling-circle synthesis of a nascent nucleic acid strand.

17. The method of claim 1, wherein the template nucleic acid is an RNA or DNA molecule.

18. The method of claim 1, wherein the enzyme is a polymerase enzyme.

19. The method of claim 7, wherein the polymerase enzyme is selected from the group consisting of a DNA polymerase, an RNA polymerase, a reverse transcriptase, and a derivative thereof.

20. The method of claim 7, wherein the polymerase enzyme is a Φ29 polymerase comprising at least one mutation at a position selected from the group consisting of K392, K422, I93, M188, K392, V399, T421, K422; S95, Y101, M102; Q99, L123, K124, T189, A190; G191, S388; P127, L384, N387, S388; and L389, Y390, and G391.

21. The method of claim 7, wherein the polymerase enzyme displays strand displacement activity.

22. The method of claim 1, wherein the processing results in the synthesis of a nascent nucleic acid strand.

23. The method of claim 1, wherein the processing is a single-molecule sequencing reaction.

24. The method of claim 22, wherein the monitoring detects incorporation of single nucleotides into the nascent nucleic acid strand to generate a sequence read that is complementary to the template nucleic acid.

25. The method of claim 24, wherein the change occurs at the modification.

26. The method of claim 24, wherein the change occurs at one or more positions upstream or downstream of the modification.

27. The method of claim 24, wherein the change comprises incorporation of a non-natural nucleotide analog at a position in the nascent nucleic acid strand that is complementary to a position in the template nucleic acid that comprises the modification.

28. The method of claim 27, wherein the change further comprises a kinetic difference in the processing.

29. The method of claim 27, wherein the modification is a methylated cytosine and the non-natural nucleotide analog is a modified guanine nucleotide analog.

30. The method of claim 27, wherein the modification is a modified base that forms a non-natural base pair with the non-natural nucleotide analog, the non-natural base pair being selected from the group consisting of isocytosine and isoguanine; 5-methylisocytosine and isoguanine; Im-N^O and Im-O^N; A* and T*; and 8-oxoG and adenine.

31. The method of claim 27, wherein the non-natural nucleotide analog comprises a detectable label that distinguishes its incorporation from incorporation of other nucleotides into the nascent nucleic acid strand.

32. The method of claim 24, wherein the single nucleotides are differentially labeled to be distinguishable from one another during the monitoring.

33. The method of claim 32, wherein the single nucleotides comprise a label linked to a phosphate group, wherein the phosphate group is removed during the processing.

34. The method of claim 24, wherein the sequence read contains multiple copies of at least a portion complementary to the template nucleic acid.

35. The method of claim 24, further comprising mapping the modification within the template nucleic acid, the mapping comprising:

f) analyzing a portion of the sequence read that was generated immediately prior to, during, or immediately after the detecting the change in processing to determine a sequence complementary to the template nucleic acid;

g) determining the complement of the sequence complementary to the template nucleic acid in f; and

h) mapping the modification at a position in the template nucleic acid that is proximal to the complement of the sequence complementary to the template nucleic acid in f.

36. The method of claim 1, wherein the change is a kinetic difference in the processing.

37. The method of claim 36, wherein the kinetic difference is detected as an alteration in interpulse duration during the processing.

38. The method of claim 36, wherein the kinetic difference is detected as an alteration in pulse width during the processing.

39. The method of claim 36, wherein the kinetic difference is indicative of what type of modification is present in the template nucleic acid.

40. The method of claim 1, wherein the monitoring occurs in real time during the processing.

41. The method of claim 1, wherein the template nucleic acid and the enzyme form a complex that is immobilized at a reaction site on a substrate.

42. The method of claim 41, wherein a plurality of complexes are immobilized at optically resolvable reaction sites on the substrate, wherein a single complex immobilized at one of the reaction sites is optically resolvable from any other of the complexes immobilized at any other of the reaction sites.

43. The method of claim 42, wherein the optically resolvable reaction sites are nanometer-scale apertures in the substrate.

44. The method of claim 43, wherein the nanometer-scale apertures are optical confinements.

45. The method of claim 44, wherein the optical confinements are zero-mode waveguides.

46. The method of claim 1, wherein the template nucleic acid is plurality of template nucleic acids that are optically resolvable from one another during the monitoring.

47. The method of claim 1, wherein the template nucleic acid is not amplified prior to the contacting.

48. The method of claim 1, wherein the modification is a secondary structure formation and the change is an alteration in one or more kinetic characteristics of the processing, the method further comprising:

f) generating a sequence read for the template nucleic acid before and after the change;

g) identifying a first portion of the sequence read generated before the change that is complementary to a second portion of the sequence read generated after the change; and

h) determining a likelihood that a secondary structure formed by association of the first portion to the second portion was present in the template nucleic acid during the processing based at least upon the nucleotide composition of the first portion and the second portion.

49. A method of detecting binding of an agent to a single nucleic acid template, comprising:

a) providing the single nucleic acid template in complex with a polymerase;

b) introducing a reaction mixture to the complex, wherein the reaction mixture comprises the agent; and

c) monitoring synthesis of a polynucleotide by the polymerase, wherein the polynucleotide is complementary to the single nucleic acid template, and wherein a change in the synthesis is indicative of binding of the agent to the single nucleic acid template.

50. The method of claim 49, wherein the agent is selected from the group consisting of a transcription factor, a polymerase, a reverse transcriptase, a histone, a restriction enzyme, an antibody, a nucleic acid binding protein, and a nucleic acid binding agent.

51. The method of claim 49, wherein the single nucleic acid template is selected from the group consisting of a double-stranded DNA, a double-stranded RNA, a single-stranded DNA, a single-stranded RNA, a DNA/RNA hybrid, and a combination thereof.

52. The method of claim 49, further comprising determining a consensus binding site of the agent, said determining comprising:

d) performing a plurality of sequencing-by-synthesis reactions on a set of single nucleic acid templates in the presence of the agent to generate a set of binding-affected nascent polynucleotide sequences;

e) performing a plurality of sequencing-by-synthesis reactions on the set of single nucleic acid templates in the absence of the agent to generate a set of full-length nascent polynucleotide sequences;

f) analyzing the binding-affected nascent polynucleotide sequences to determine a location at which the agent bound the single nucleic acid template during the sequencing-by-synthesis reactions in the presence of the agent; and

g) identifying a sequence common to the full-length nascent polynucleotide sequences at the location determined in (iii), thereby identifying the consensus binding site of the agent.

53. The method of claim 52, wherein the binding-affected nascent polynucleotide sequences are truncated nascent polynucleotide sequences.

54. The method of claim 52, wherein the binding-affected nascent polynucleotide sequences are nascent polynucleotide sequences whose synthesis was paused at the location determined in f.

55. A method for detecting a modification in a single nucleic acid template during a sequencing-by-synthesis reaction, the method comprising:

a) providing the single nucleic acid template in complex with a polymerase;

b) introducing a reaction mixture to the complex, wherein the reaction mixture comprises an agent that specifically binds to the modification; and

c) monitoring synthesis of a polynucleotide by the polymerase, wherein the polynucleotide is complementary to the single nucleic acid template, and wherein a pause or cessation of the synthesis of the polynucleotide is indicative of binding of the agent to the single nucleic acid template, thereby detecting the modification in the single nucleic acid template.

56. The method of claim 55, wherein the modification is an 8-oxoG lesion.

57. The method of claim 55, wherein the agent is a protein is selected from the group consisting of hOGG1, FPG, yOGG1, AlkA, Nth, Nei, MutY, UDG, SMUG, TDG, NEIL, an antibody against 8-oxoG, or a binding domain thereof.

58. The method of claim 55, wherein the modification is a methylated base.

59. The method of claim 58, wherein the agent is a protein selected from the group consisting of MECP2, MBD1, MBD2, MBD4, UHRF1, an antibody against the methylated base, or a binding domain thereof.

60. The method of claim 55, wherein the modification is a secondary structure formation in the nucleic acid template.

61. The method of claim 55, wherein the complex is immobilized in an optical confinement.

62. The method of claim 55, wherein the template is selected from the group consisting of a single-stranded linear nucleic acid template, a single-stranded circular nucleic acid template, double-stranded linear nucleic acid template, a double-stranded circular nucleic acid template, and a combination thereof.

63. The method of claim 55, further comprising repairing the modification by including components of damage repair machinery in the reaction mixture.

64. The method of claim 63, wherein the readlength of the sequencing-by-synthesis reaction is longer than that for a further sequencing-by-synthesis reaction performed with the single nucleic acid template in complex with the polymerase, but absent the agent and the damage repair machinery.

65. A method for bypassing one or more modifications in a single nucleic acid template during a sequencing-by-synthesis reaction, the method comprising:

a) providing the single nucleic acid template in complex with a sequencing engine;

b) introducing a reaction mixture to the complex, wherein the reaction mixture comprises a bypass polymerase;

c) initiating the sequencing-by-synthesis reaction;

d) monitoring synthesis of a polynucleotide by the sequencing engine, wherein the polynucleotide is complementary to the single nucleic acid template, and wherein a pause or cessation of the synthesis of the polynucleotide is indicative that the sequencing engine has encountered a modification in the single nucleic acid template;

e) subsequently monitoring synthesis of the polynucleotide by the bypass polymerase, which is indicative that the modification is being bypassed; and

f) repeating steps c) and d) each time a further modification is encountered in the single nucleic acid template, thereby bypassing one or more modifications in a single nucleic acid template during a sequencing-by-synthesis reaction.

66. The method of claim 65, wherein the bypass polymerase comprises a detectable label and detection of a signal from the detectable label during the sequencing-by-synthesis reaction is indicative that the bypass polymerase is actively synthesizing the polynucleotide.

67. The method of claim 65, wherein the readlength of the sequencing-by-synthesis reaction is longer than that for a further sequencing-by-synthesis reaction performed with the single nucleic acid template in complex with the sequencing engine, but absent the bypass polymerase.

68. The method of claim 65, wherein the reaction mixture comprises multiple different bypass polymerases and a processivity factor.

69. The method of claim 65, wherein at least one of the single nucleic acid template, the sequencing engine, and the bypass polymerase is immobilized, directly or indirectly, in an optical confinement.

70. The method of claim 69, wherein the single nucleic acid template is immobilized by hybridization to an oligonucleotide primer immobilized in the optical confinement.

71. The method of claim 65, wherein the single nucleic acid template is processed by the sequencing engine multiple times at a single reaction site, and further wherein redundant sequence data is generated during the monitoring of step d).

72. The method of claim 71, wherein the single nucleic acid template is a circular nucleic acid template.

73. The method of claim 65, wherein the single nucleic acid template comprises multiple copies of a nucleic acid segment of interest, and the sequencing-by-synthesis reaction generates a polynucleotide comprising multiple copies of a segment complementary to the segment of interest, and further wherein redundant sequence data is generated during the monitoring of step d).

74. A composition comprising:

a) a substrate comprising a reaction site that is optically resolvable form any other reaction site on the substrate;

b) a single complex immobilized at the reaction site, wherein the complex comprises a template nucleic acid and a sequencing engine;

c) a mixture of incorporatable nucleotides or nucleotide analogs; and

d) at least one modification in the template nucleic acid, wherein the template at or proximal to the modification is processed differently than the template distal from the modification.

75. The composition of claim 74, wherein the modification is a non-natural base in the template nucleic acid.

76. The composition of claim 74, wherein the modification is in a strand of the template nucleic acid that is complementary to a nascent strand synthesized by the sequencing engine.

77. The composition of claim 74, wherein the modification is in a strand of the template nucleic acid that is displaced by the sequencing engine.

78. The composition of claim 74, further comprising at least one type of non-incorporatable nucleotide analog.

79. The composition of claim 74, wherein the mixture of incorporatable nucleotides or nucleotide analogs comprises at least one type of non-natural nucleotide analog.

80. The composition of claim 74, wherein the mixture of incorporatable nucleotides or nucleotide analogs comprises different types of nucleotides or nucleotide analogs, and further wherein the different types are distinguishable from one another because each type carries a distinct detectable label.

81. The composition of claim 74, further comprising an agent other than the sequencing engine that binds to the modification.

82. The composition of claim 74, further comprising an agent that chemically or enzymatically alters the modification.

83. The composition of claim 74, further comprising a nascent strand generated by the sequencing engine, wherein the nascent strand is complementary to the template nucleic acid.

84. The composition of claim 83, wherein the nascent strand comprises multiple copies of regions complementary to the template nucleic acid.

85. The composition of claim 74, further comprising a nanometer-scale aperture in the substrate, wherein the reaction site is disposed within the nanometer-scale aperture.

86. The composition of claim 74, wherein the template nucleic acid comprises internally complementary regions, and further wherein the modification is located within one of the internally complementary regions.

87. The composition of claim 85, wherein the nanometer-scale aperture is a ZMW.

88. A system for identification of modifications within a nucleic acid template, the method comprising:

a) a solid support having a polymerase complex disposed thereon, the polymerase complex comprising a nucleic acid template comprising a modification;

b) a mounting stage configured to receive the solid support;

c) an optical train positioned to be in optical communication with at least a portion of the solid support to detect signals emanating therefrom;

d) a translation system operably coupled to the mounting stage or the optical train for moving one of the optical train and the solid support relative to the other; and

e) a data processing system operably coupled to the optical train.

89. The system of claim 88, wherein the polymerase complex comprises a polymerase enzyme that is actively processing the nucleic acid template.

90. The system of claim 89, wherein the polymerase complex comprises a polymerase enzyme that is processively synthesizing a nascent strand by template-directed synthesis.

91. The system of claim 89, wherein the optical train detects signals emanating from the solid support during the processing of the nucleic acid template.

92. The system of claim 88, wherein the solid support comprises at least one zero-mode waveguide.

93. A machine-implemented method for transforming reaction data into modification detection data, wherein the reaction data is representative of a series of events during a sequencing-by-synthesis reaction wherein a nascent strand is synthesized based upon a nucleotide sequence of a template nucleic acid, and the modification detection data is representative of a presence of one or more modifications within a template nucleic acid, the machine-implemented method comprising:

a) creation of a classifier to distinguish between true incorporations and stochastic pulses in an observed signal trace that comprises the true incorporations and the stochastic pulses;

b) slide a fixed-length window across the observed signal trace and use the classifier to count the true incorporations and the stochastic pulses; and

c) based upon the results of step b, identify regions of the observed signal trace in which a density of the stochastic pulses exceeds a density of the true incorporations, wherein the regions are characterized as candidate regions comprising at least one of the one or more modifications.

94. The method of claim 93, wherein steps a-c are performed via a user interface implemented in a machine that comprises instructions stored in machine-readable medium and a processor that executes the instructions.

95. A computer program product comprising a computer usable medium having

a) a computer readable program code embodied therein, said computer readable program code adapted to be executed to implement the method of claim 93; and

b) the machine-readable medium on which the results of steps a-c are stored.

96. A machine-implemented method for transforming reaction data into modification detection data, wherein the reaction data is representative of a series of events during a sequencing-by-synthesis reaction wherein a nascent strand is synthesized based upon a nucleotide sequence of a template nucleic acid, and the modification detection data is representative of a presence of one or more modifications within a template nucleic acid, the machine-implemented method comprising:

a) creation of a classifier to distinguish between true incorporations and stochastic pulses in an observed signal trace that comprises the true incorporations and the stochastic pulses;

b) slide a fixed-length window across the observed signal trace and determine times between the true incorporations; and

c) based upon the results of step b, identify regions of the observed signal trace that have a large difference in time between the true incorporations, wherein the regions are characterized as candidate regions comprising at least one of the one or more modifications.

97. The method of claim 96, wherein steps a-c are performed via a user interface implemented in a machine that comprises instructions stored in machine-readable medium and a processor that executes the instructions.

98. A computer program product comprising a computer usable medium having

a) a computer readable program code embodied therein, said computer readable program code adapted to be executed to implement the method of claim 96; and

b) the machine-readable medium on which the results of steps a-c are stored.

99. A machine-implemented method for transforming reaction data into modification detection data, wherein the reaction data is representative of a series of events during a sequencing-by-synthesis reaction wherein a nascent strand is synthesized based upon a nucleotide sequence of a template nucleic acid, and the modification detection data is representative of a presence of one or more modifications within a template nucleic acid, the

machine-implemented method comprising a segmenting algorithm based on a hidden Markov model architecture.

100. A machine-implemented method for transforming reaction data into modification detection data, wherein the reaction data is representative of a series of events during a sequencing-by-synthesis reaction wherein a nascent strand is synthesized based upon a nucleotide sequence of a template nucleic acid, and the modification detection data is representative of a presence of one or more modifications within a template nucleic acid, the machine-implemented method comprising a segmenting algorithm based on a conditional random field framework.

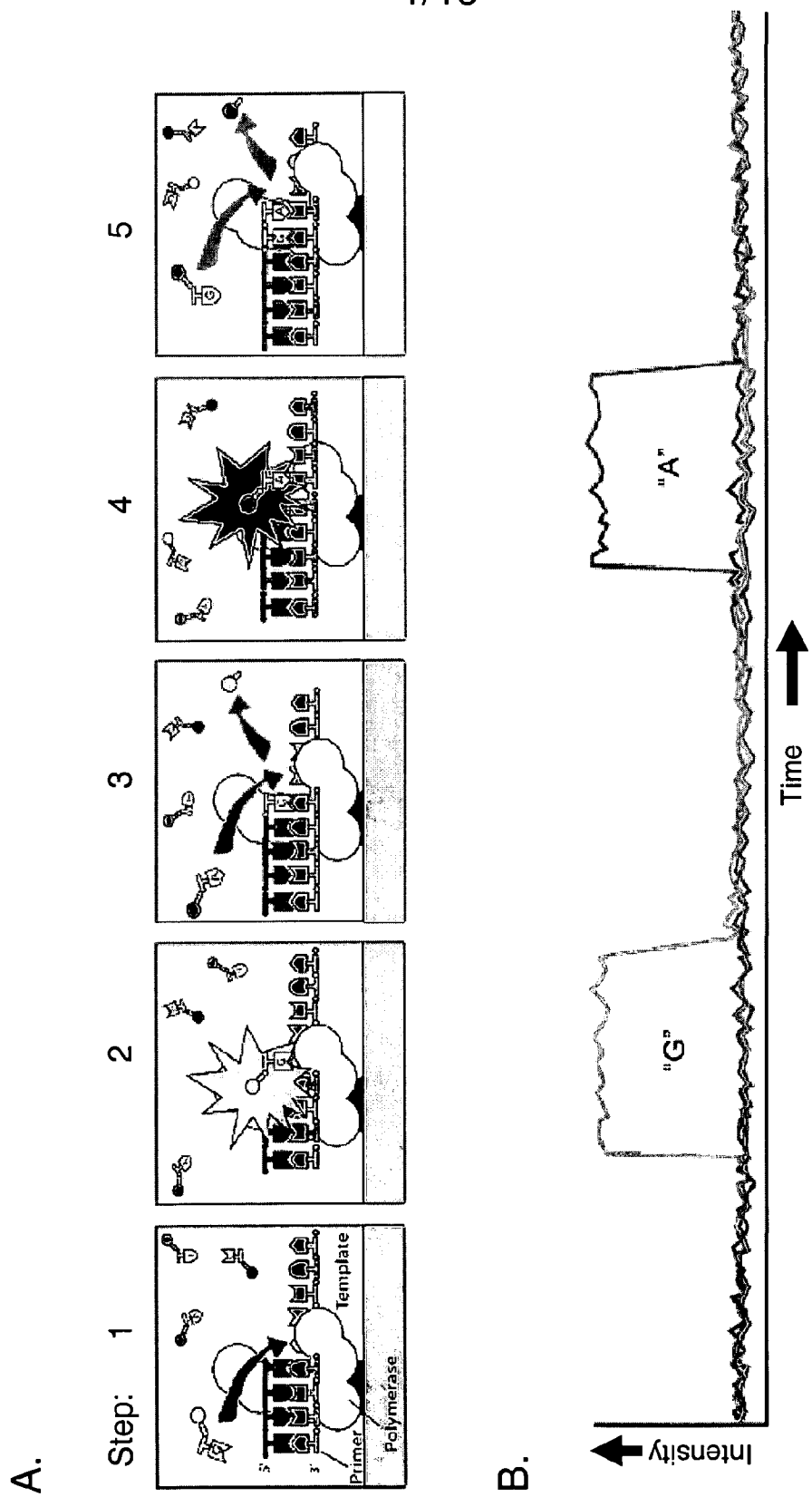

A.
Step: 1
2
3
4
5
Template
Primer
Polymerase
B.
"G"
"A"
Intensity
Time

Figure 1

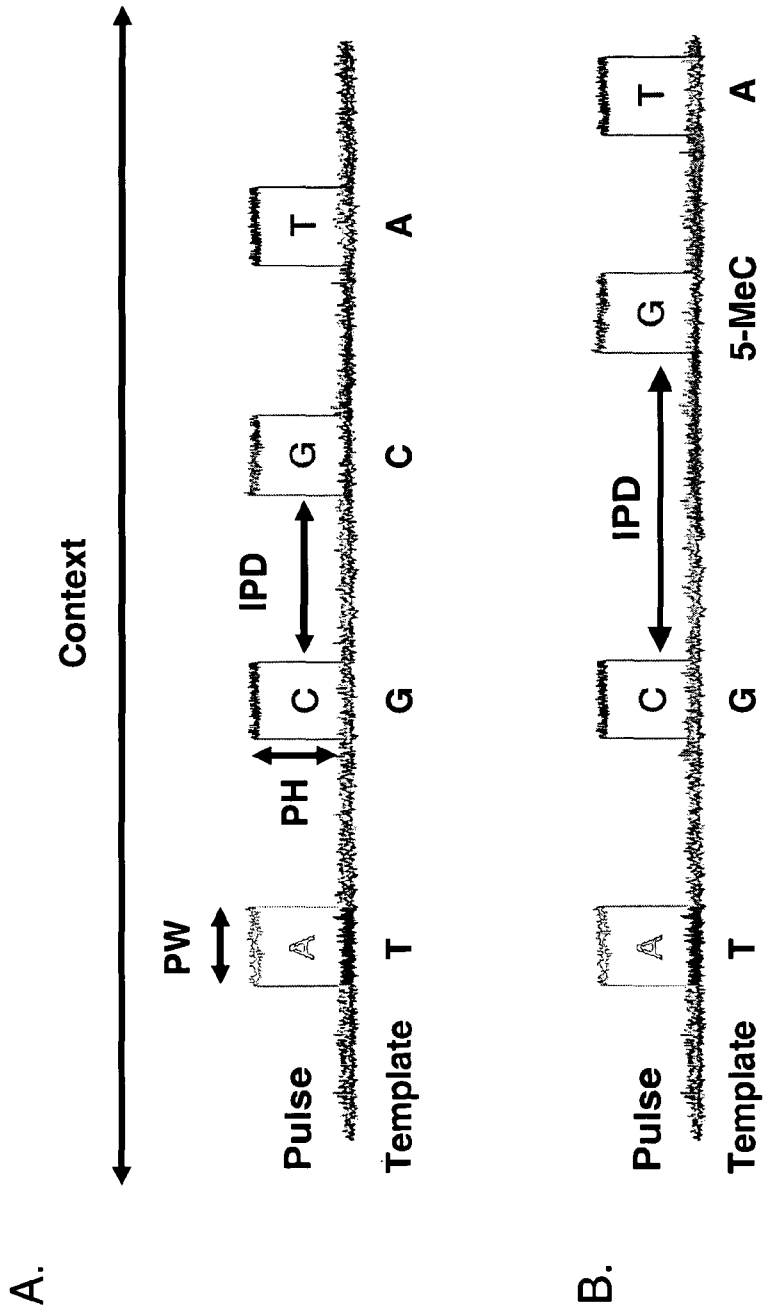

A.
Context
PW
PH
IPD
Pulse
Template
T
A
C
G
T
G
C
A
B.
IPD
Pulse
Template
T
A
C
G
T
G
5-MeC
A

Figure 2

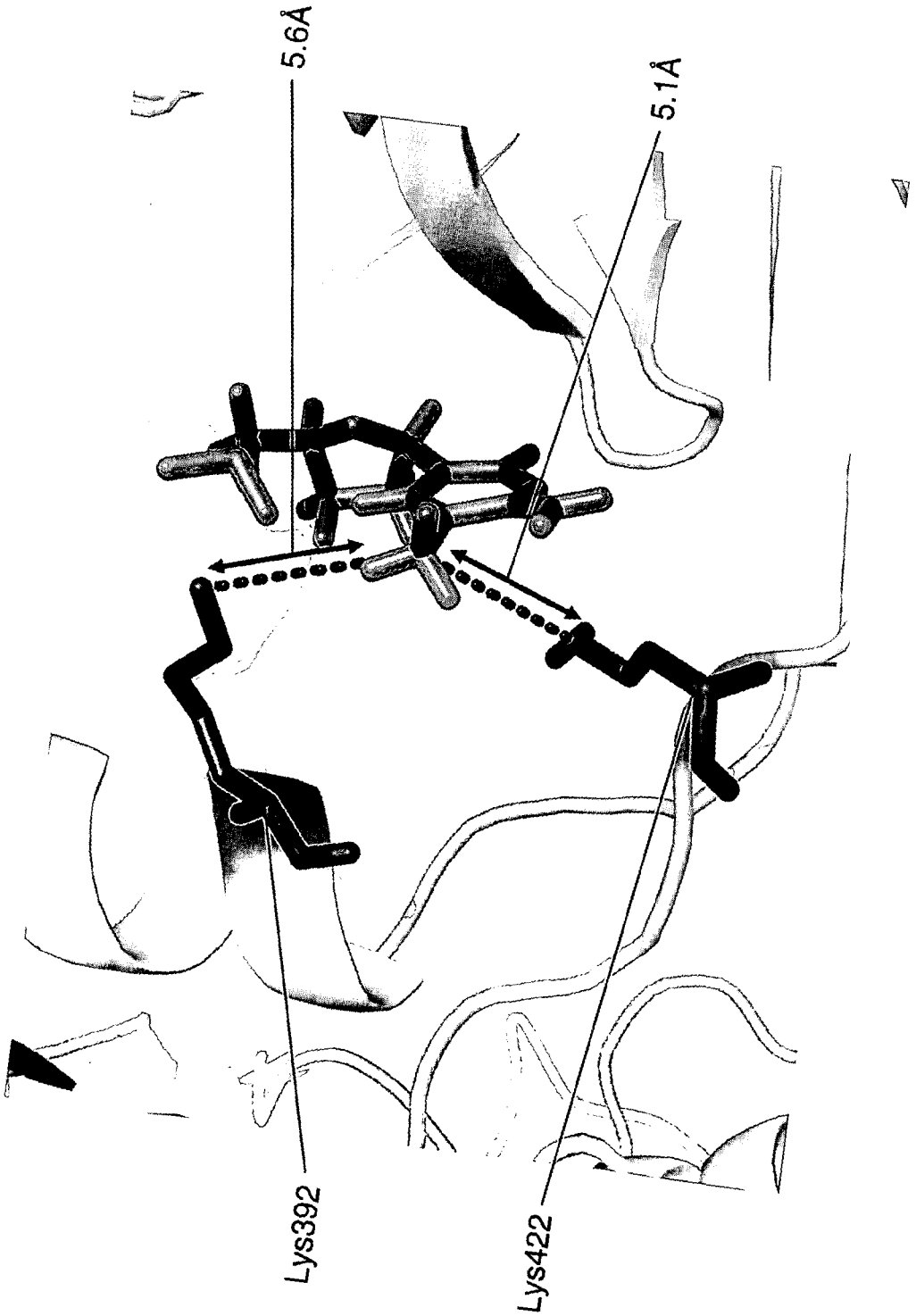


Figure 3

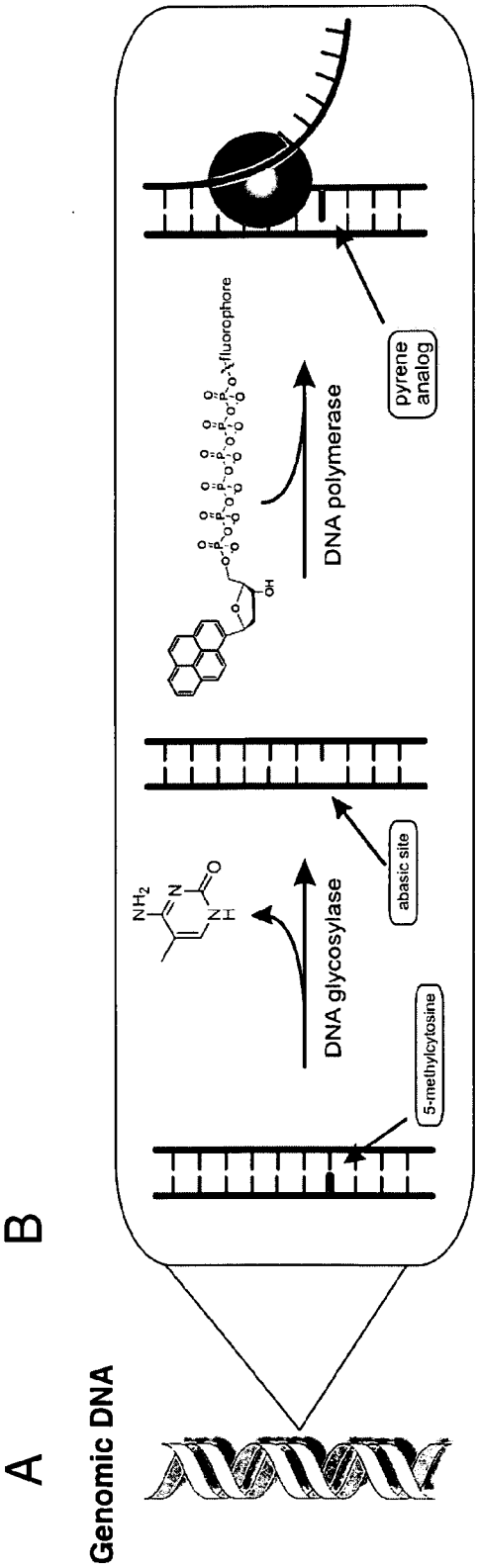

A
Genomic DNA
B
5-methylcytosine
DNA glycosylase
abasic site
DNA polymerase
pyrene analog
fluorophore

Figure 4

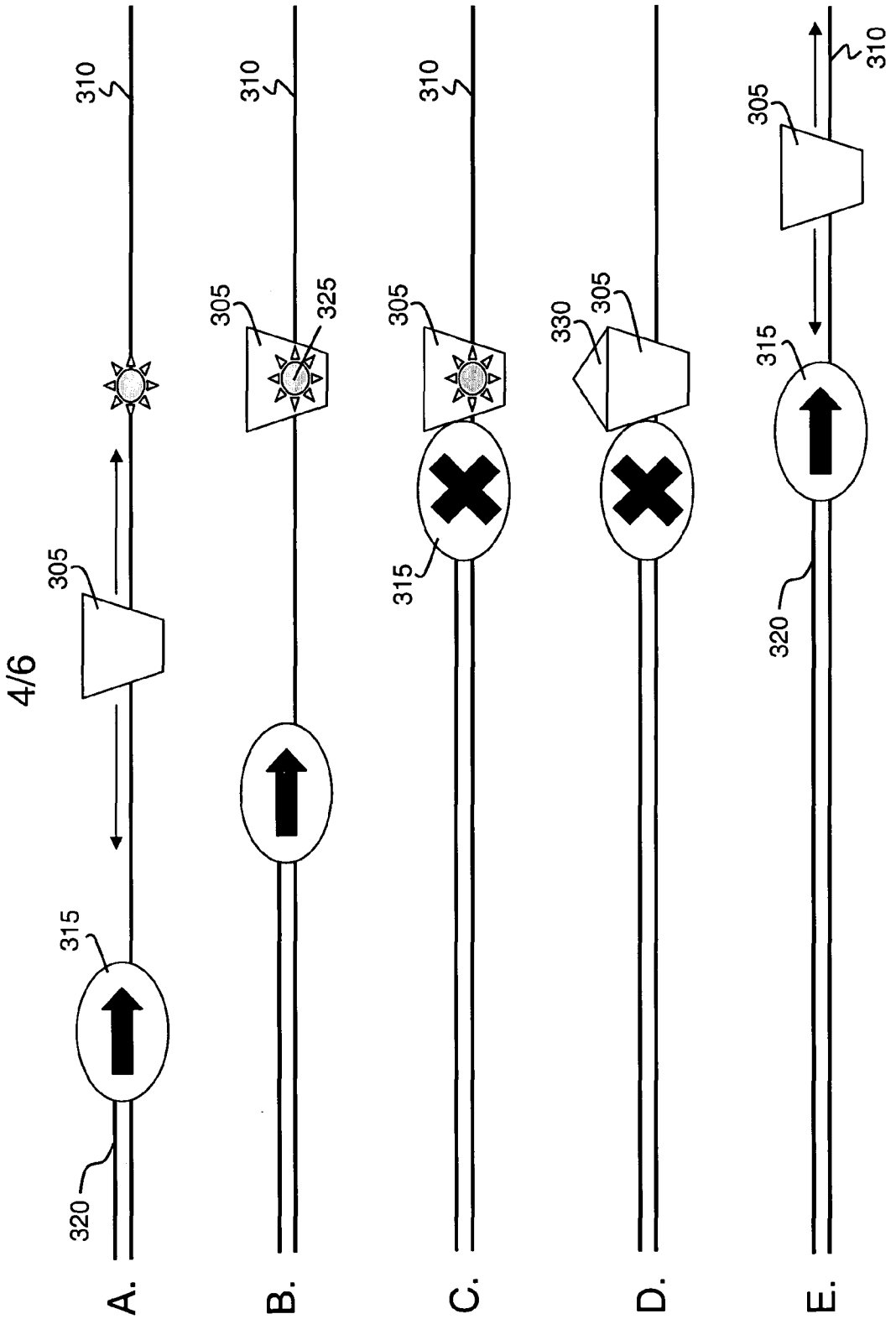


Figure 5

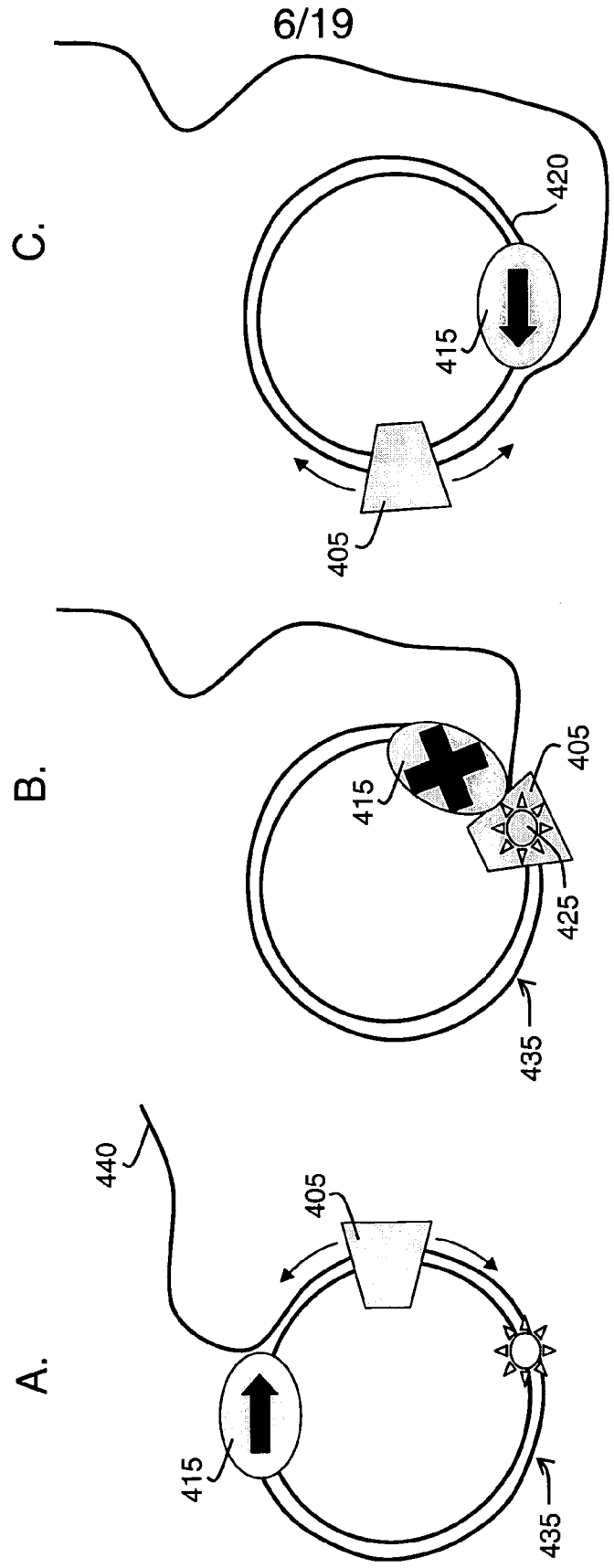


Figure 6

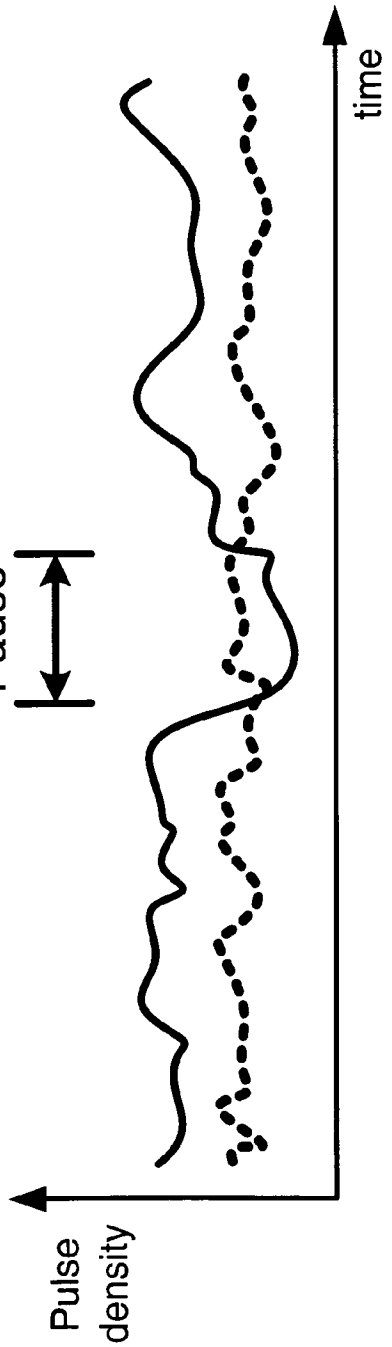


Figure 7

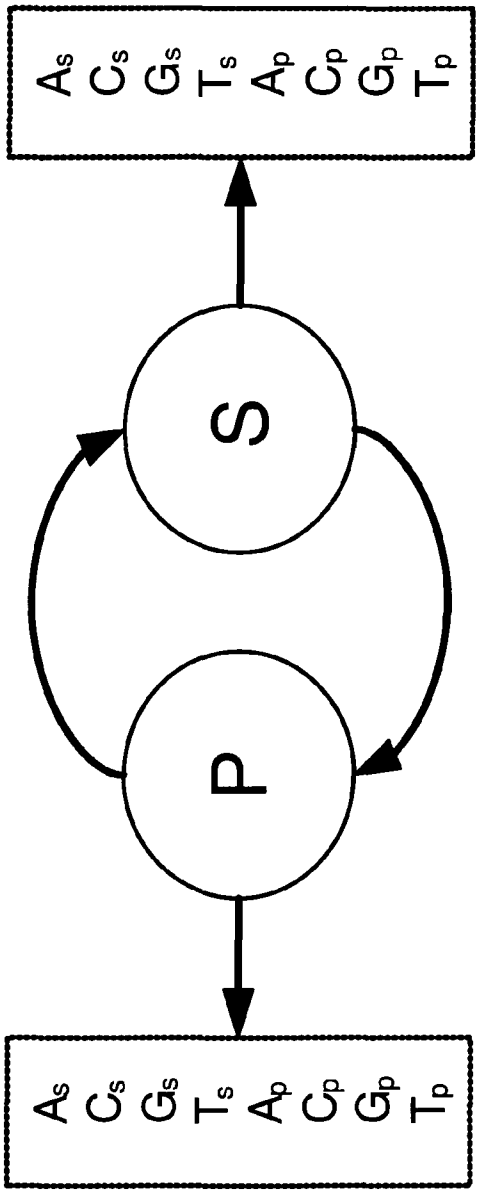


Figure 8

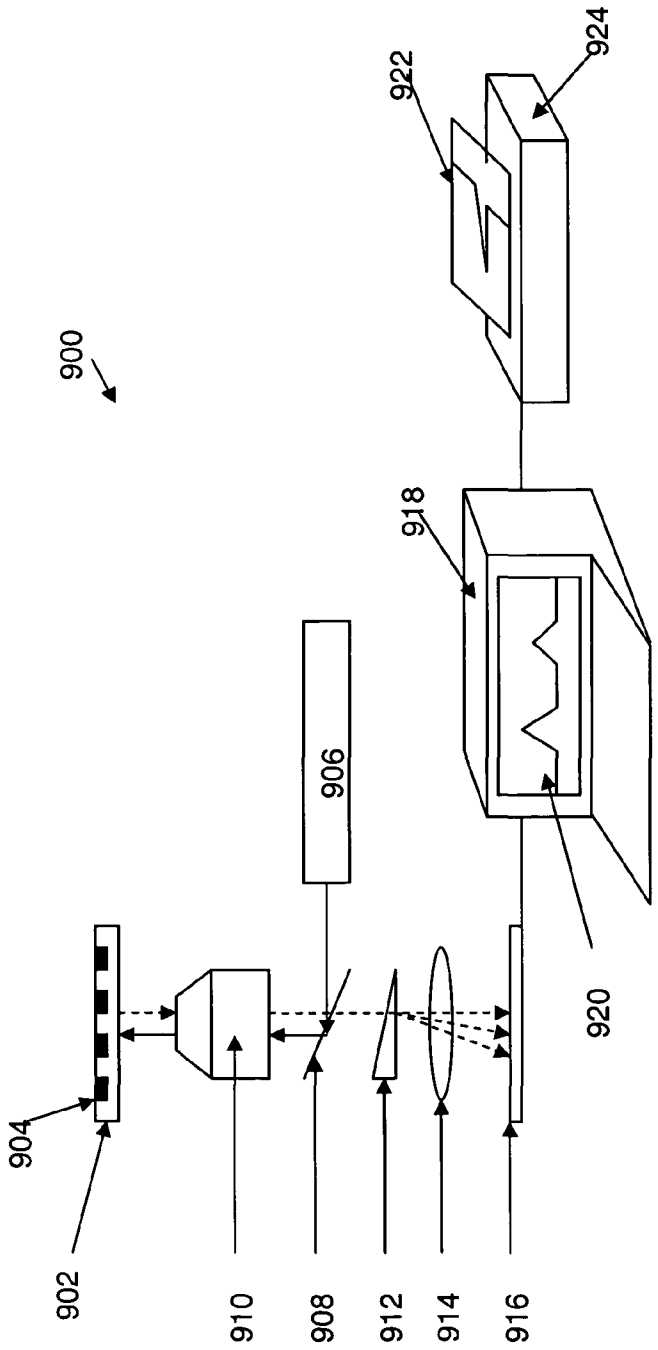


Figure 9

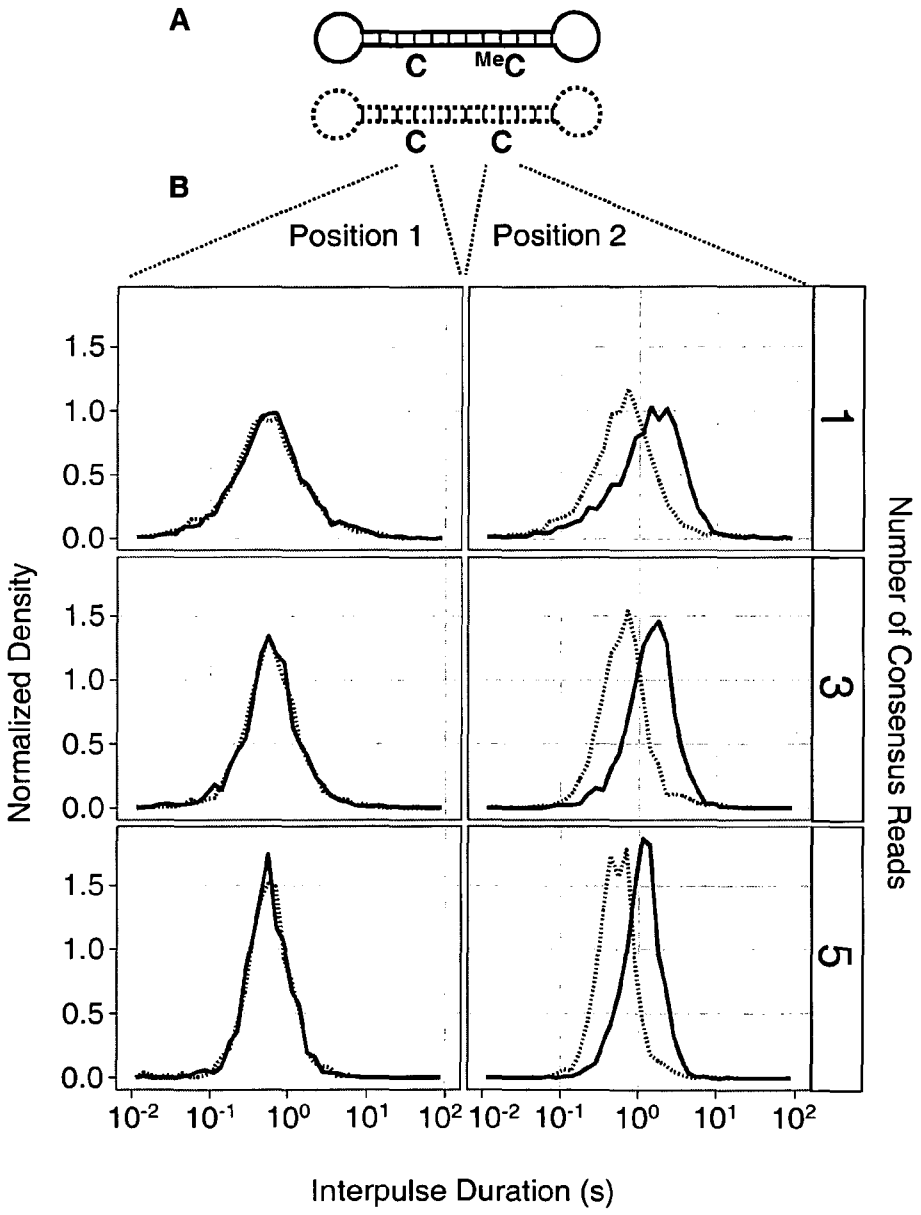

A
C
MeC
C
C
B
Position 1
Position 2
1.5
1.0
0.5
0.0
1
3
5
Normalized Density
Number of Consensus Reads
10-2 10-1 100 101 102
Interpulse Duration (s)

Figure 10

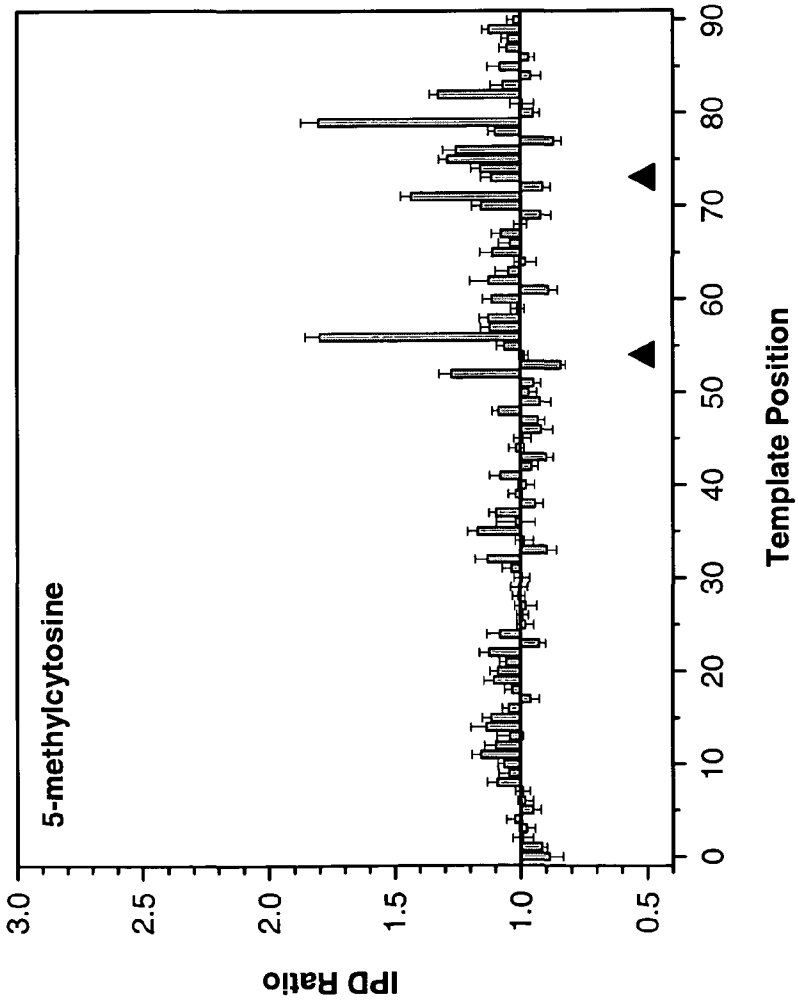

5-methylcytosine
IPD Ratio
3.0
2.5
2.0
1.5
1.0
0.5
0
10
20
30
40
50
60
70
80
90
Template Position

Figure 11

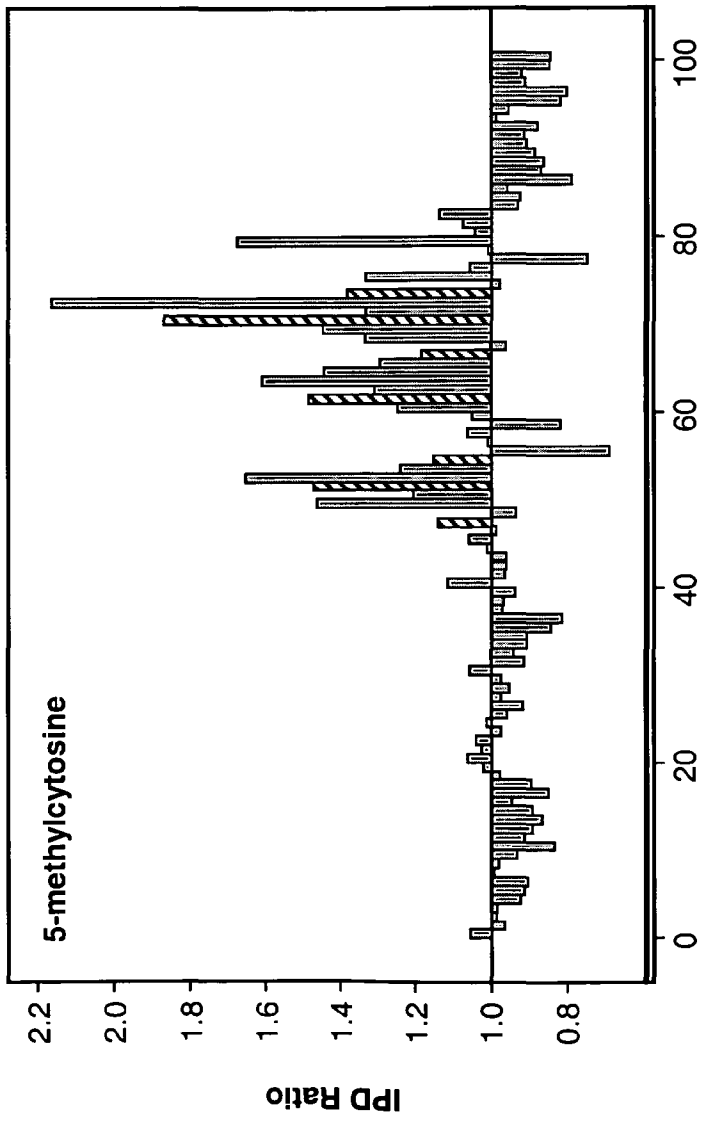

5-methylcytosine
IPD Ratio
2.2
2.0
1.8
1.6
1.4
1.2
1.0
0.8
0
20
40
60
80
100
Template Position

Figure 12

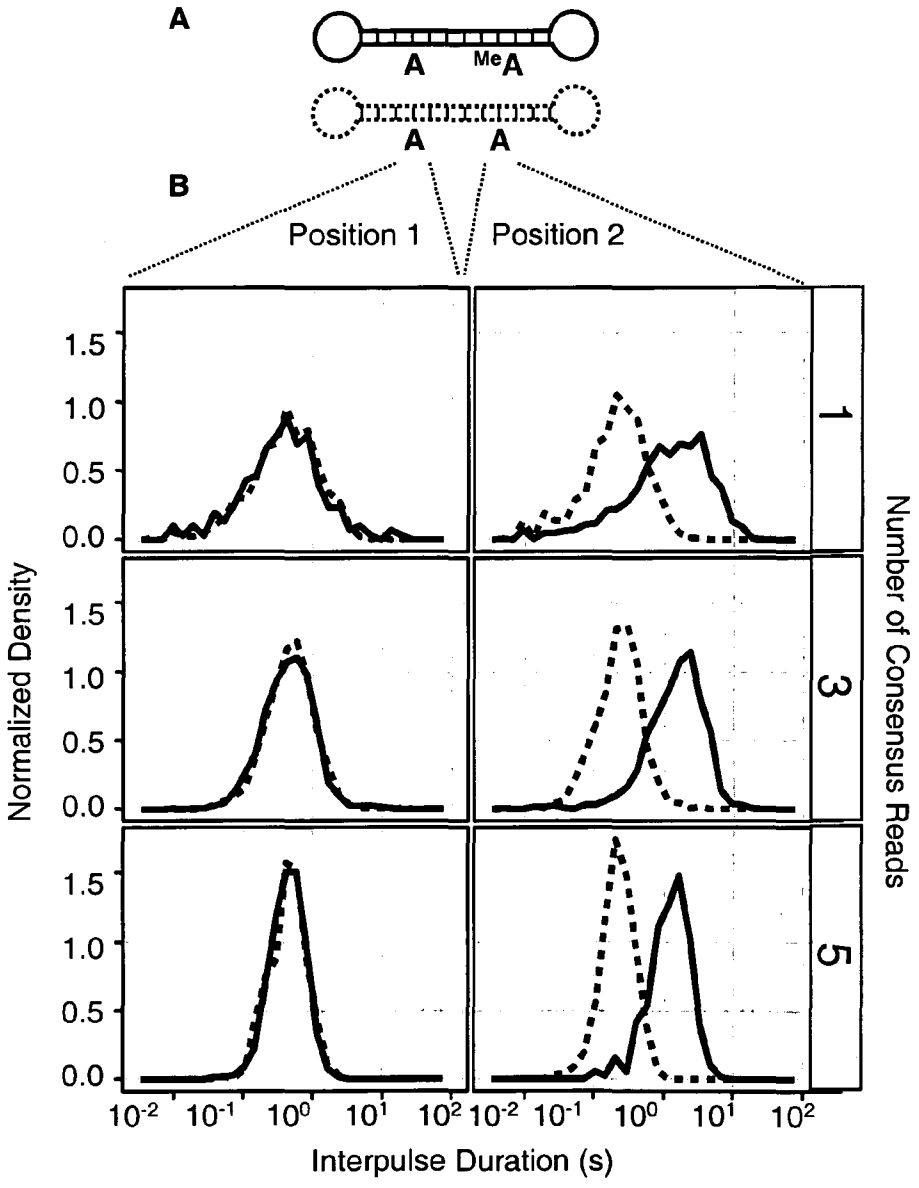

A
A
Me A
A
A
B
Position 1
Position 2
1.5
1.0
0.5
0.0
1
3
5
Normalized Density
Number of Consensus Reads
10^{-2} 10^{-1} 10^{0} 10^{1} 10^{2}
Interpulse Duration (s)

Figure 13

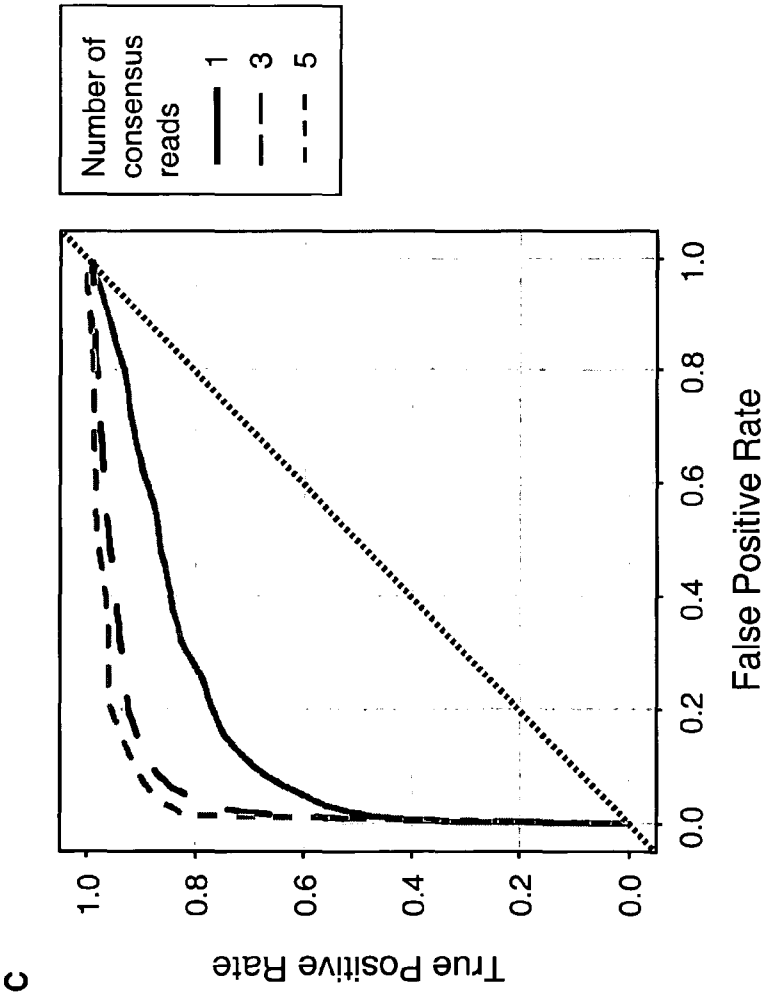


Figure 13, cont.

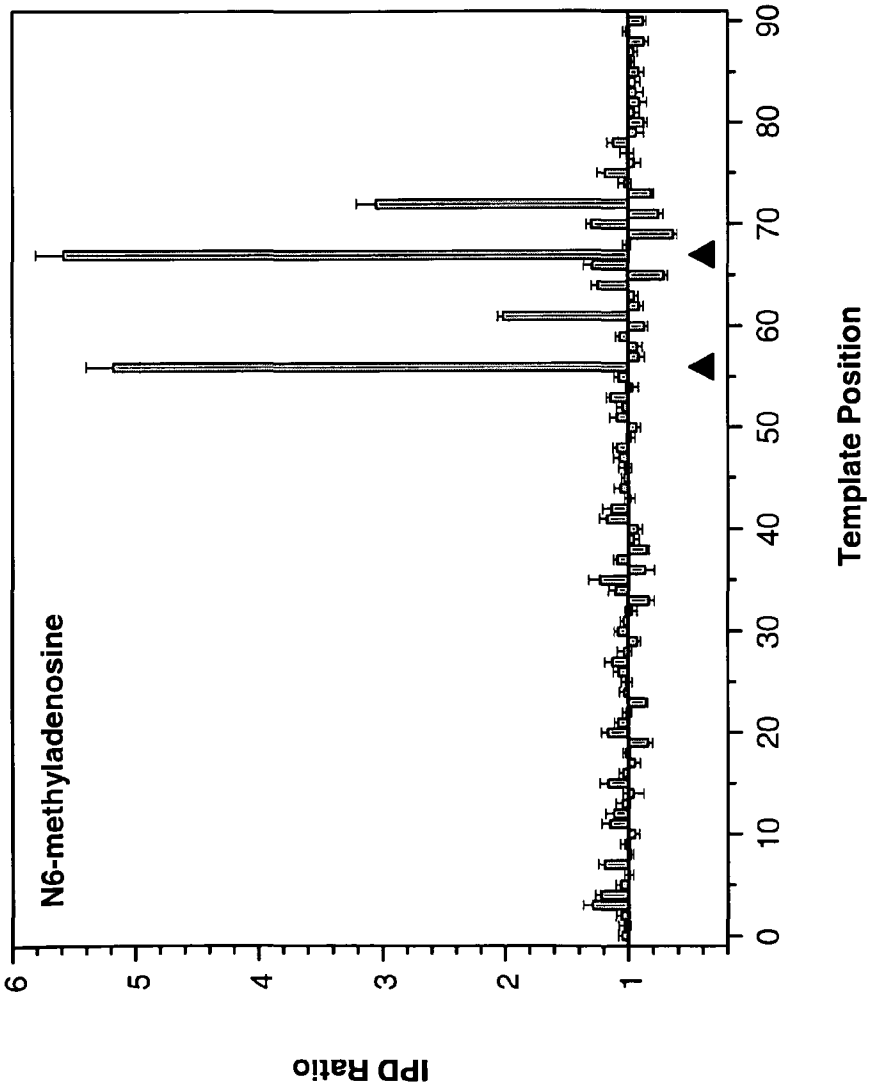

N6-methyladenosine
IPD Ratio
Template Position
1
2
3
4
5
6
0
10
20
30
40
50
60
70
80
90

Figure 14

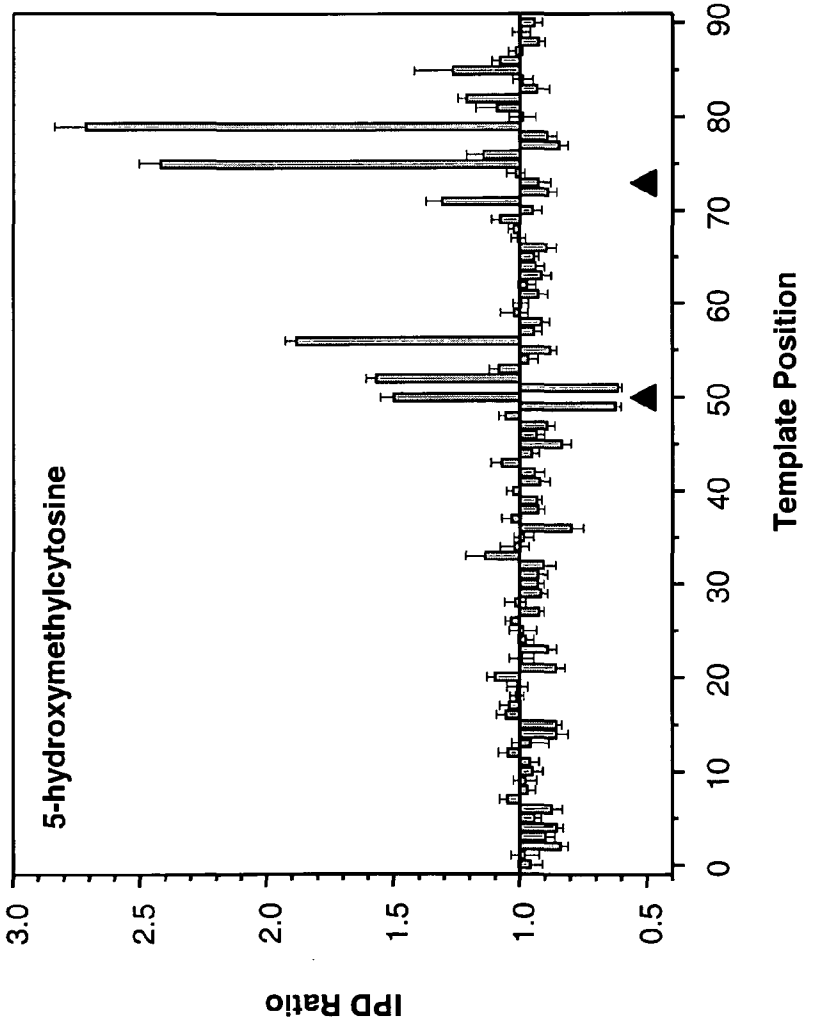

5-hydroxymethylcytosine
IPD Ratio
3.0
2.5
2.0
1.5
1.0
0.5
0
10
20
30
40
50
60
70
80
90
Template Position

Figure 15

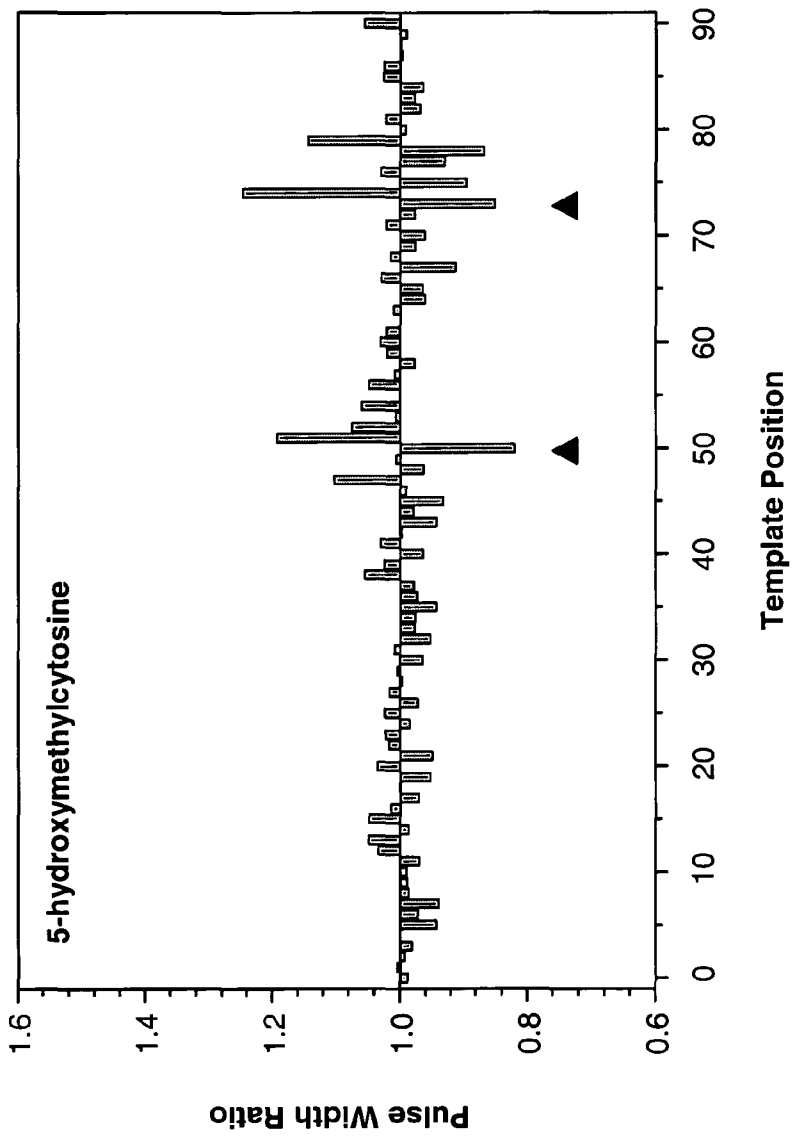

5-hydroxymethylcytosine
Pulse Width Ratio
1.6
1.4
1.2
1.0
0.8
0.6
0
10
20
30
40
50
60
70
80
90
Template Position

Figure 16

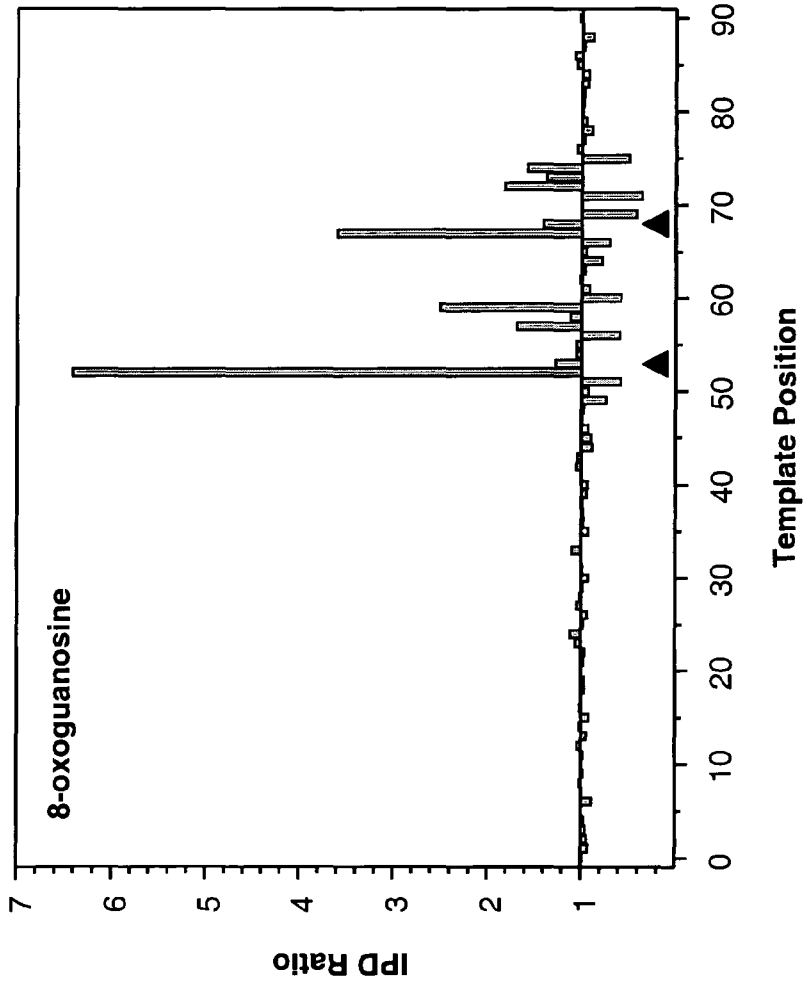

8-oxoguanosine
IPD Ratio
Template Position
1
2
3
4
5
6
7
0
10
20
30
40
50
60
70
80
90

Figure 17

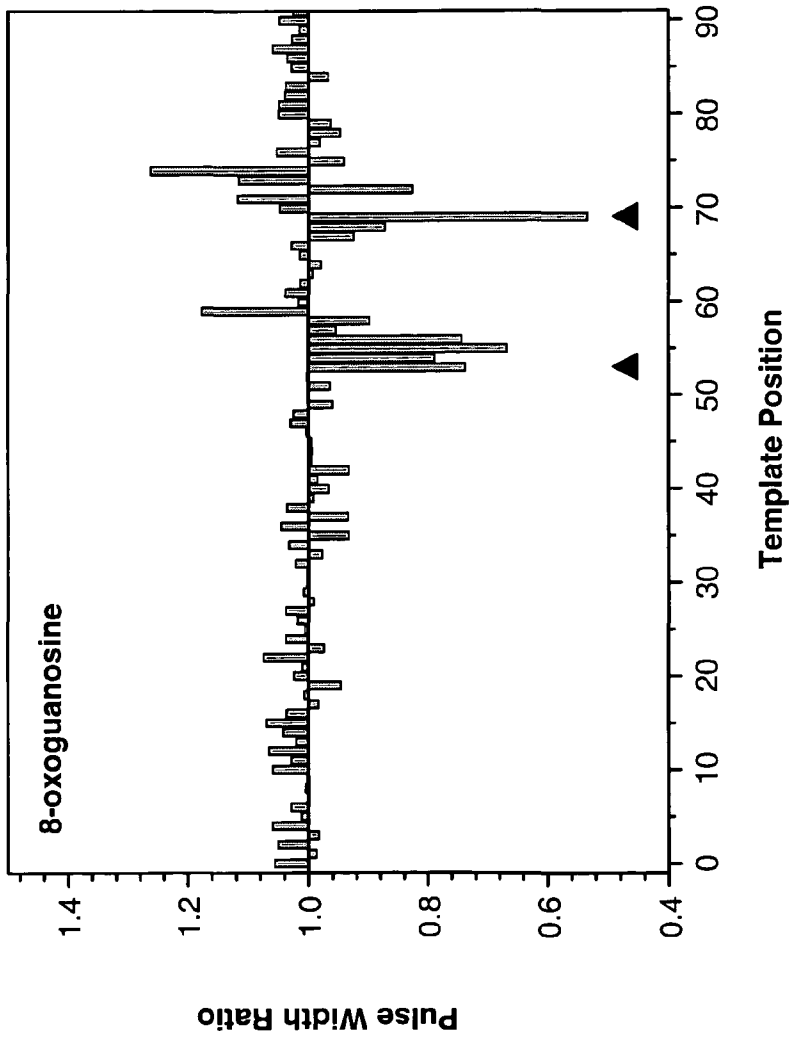

8-oxoguanosine
Pulse Width Ratio
1.4
1.2
1.0
0.8
0.6
0.4
0
10
20
30
40
50
60
70
80
90
Template Position

Figure 18

(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
11 March 2010 (11.03.2010)

PCT

(10) International Publication Number
WO 2010/027484 A2

(51) International Patent Classification:
C12Q 1/68 (2006.01) *C12Q 1/25* (2006.01)
C12N 15/11 (2006.01)

(21) International Application Number:
PCT/US2009/004993

(22) International Filing Date:
4 September 2009 (04.09.2009)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
61/094,843 5 September 2008 (05.09.2008) US
12/370,472 12 February 2009 (12.02.2009) US

(71) Applicant *(for all designated States except US)*: **PACIFIC BIOSCIENCES OF CALIFORNIA, INC.** [US/US]; 1505 Adams Drive, Menlo Park, CA 94025 (US).

(72) Inventors; and

(75) Inventors/Applicants *(for US only)*: **PATEL, Pranav** [IN/US]; 34203 Oneil Terrace, Fremont, CA 94555 (US). **KORLACH, Jonas** [DE/US]; 36100 Toulouse Street, Newark, CA 94560 (US). **BIBILLO, Arkadiusz** [PL/US]; 10100 Torre Ave, Apt. 144, Cupertino, CA 95014 (US). **BJORNSON, Keith** [US/US]; 8098 Peachtree Avenue, Newark, CA 94560 (US). **HANES, Jeremiah** [US/US]; 3243 Spring Street, Redwood City, CA 94063 (US).

(74) Agents: **LITTLEPAGE, Paul** et al.; Quine Intellectual Property Law Group, P.C., P.O. Box 458, Alameda, CA 94501 (US).

(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IS, JP, KE, KG, KM, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PE, PG, PH, PL, PT, RO, RS, RU, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(84) Designated States *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, GH, GM, KE, LS, MW, MZ, NA, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Published:

— *without international search report and to be republished upon receipt of that report (Rule 48.2(g))*

(54) Title: ENGINEERING POLYMERASES AND REACTION CONDITIONS FOR MODIFIED INCORPORATION PROPERTIES

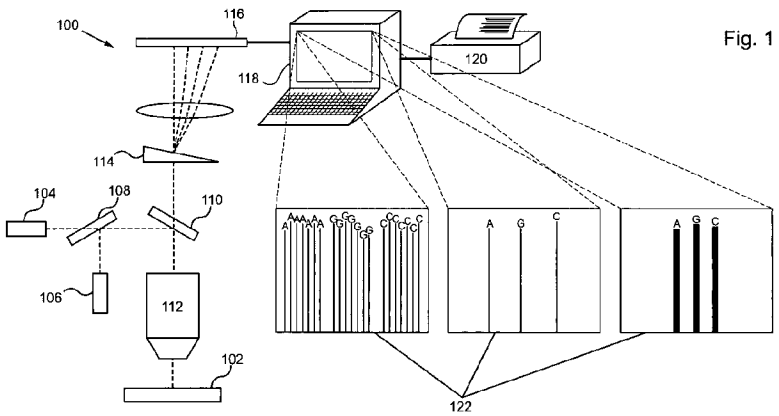


Fig. 1

(57) Abstract: Provided are methods for enhanced sequencing of nucleic acid templates. Also provided are reaction conditions that increase branching fractions during polymerization reactions. Also provided are compositions comprising modified recombinant polymerases that exhibit branching fractions that are higher than the branching fractions of the polymerases from which they were derived. Provided are compositions comprising modified recombinant polymerases that exhibit delayed translocation relative to the polymerases from which they were derived. Also provided are compositions comprising modified recombinant polymerases that exhibit increased nucleotide or nucleotide analog residence time at an active site of the polymerase. Provided are methods for generating polymerases with the aforementioned phenotypes and methods of using such polymerases to sequence a DNA template or make a DNA. Also provided are methods and nucleic acid sequencing systems for determining which labeled nucleotide is incorporated at a site during a template-dependent polymerization reaction.

WO 2010/027484 A2

ENGINEERING POLYMERASES AND REACTION CONDITIONS FOR MODIFIED INCORPORATION PROPERTIES

CROSS-REFERENCE TO RELATED APPLICATIONS

[0001] This application claims priority to Provisional U.S. Patent Application No. 61/094,843, filed September 5, 2008, and U.S. Patent Application No. 12/370,472, filed February 12, 2009, the full disclosures of which are incorporated herein by reference in their entirety for all purposes.

FIELD OF THE INVENTION

[0002] The invention is in the field of nucleic acid sequencing, for example, single molecule sequencing.

BACKGROUND OF THE INVENTION

[0003] High throughput sequencing has become a central tool in the field of biotechnology and is revolutionizing personalized medicine. Many diseases and/or disorders are genetic in origin. Acquiring the genomic sequence of individual patients in a comprehensive, rapid and cost-effective manner enhances the ability of medical professionals to diagnose diseases or identify predispositions to diseases or other genetic-based disorders. Genomic sequence information also enhances the treatment of diseases by providing doctors with information regarding the efficacy of a given therapy for a particular individual.

[0004] One approach aimed at efficiently obtaining the complete genomic sequence of an organism is sequencing by incorporation, where the identity of the sequence of nucleotides in a template nucleic acid polymer is determined by identifying each complementary base that is added to a nascent strand being synthesized against the template sequence, as such bases are added. While detection of added bases may be a result of detecting a byproduct of the synthesis or extension reaction, e.g., detecting released pyrophosphate, in many systems and processes, added bases are labeled with fluorescent dyes that permit their detection. By uniquely labeling each base with a distinguishable fluorescent dye, one attaches a distinctive

detectable characteristic to each dye that is incorporated, and as a result provides a basis for identification of an incorporated base, and by extension, its complementary base upon the template sequence.

[0005] During sequencing by incorporation, nucleotide (or nucleotide analog) incorporation events are detected in real-time as the bases are incorporated into the extension product. This can be accomplished by immobilizing the complex within an optically confined space or otherwise resolved as an individual molecular complex. Some sequencing by incorporation methods employ nucleotide analogs that include fluorescent labels coupled to the polyphosphate chain of the analog, which are then exposed to the complex. Upon incorporation, the nucleotide – along with its fluorescent label – is retained by the complex for a time and in a manner that permits the detection of a signal "pulse" from the fluorescent label at the incorporation site. Upon completion of incorporation, all but the alpha phosphate group of the nucleotide is cleaved away, liberating the label from retention by the complex, and diffusing the signal from that label.

[0006] Thus, during an incorporation event, a complementary nucleotide analog, including its fluorescent label, is effectively "immobilized" for a time at the incorporation site, and the fluorescent label is subsequently released and diffuses away when incorporation is completed. Detecting the localized "pulses" of fluorescent tags immobilized at the incorporation site, and distinguishing those pulses from a variety of other signals and background noise, allows bases to be called in real-time as they are incorporated. Further details regarding base calling during sequencing by incorporation methods are found in Tomaney et al. PCT Application Serial No. PCT/US2008/065996 METHODS AND PROCESSES FOR CALLING BASES IN SEQUENCING BY INCORPORATION METHODS, incorporated herein by reference in its entirety for all purposes.

[0007] Current real-time sequencing by incorporation methods may exhibit sub-optimal reliability and accuracy due to missed signal pulses that contribute as errors in sequencing reads. Missed pulses derive from, e.g., insufficient residence time of the analogs at an active site of the polymerase or unlabeled or broken-fluorophore nucleotide analogs. Compositions

and methods for improving the reliability and accuracy of sequencing by incorporation are desirable.

SUMMARY OF THE INVENTION

[0008] Altered reaction conditions and modified DNA polymerases can find use in such applications as, e.g., single-molecule sequencing (SMS), genotyping analyses such as SNP genotyping using single-base extension methods, and real-time monitoring of amplification, e.g., real time PCR. The invention provides methods of sequencing a nucleic acid template, which methods utilize signal pulses or signatures from branch fraction nonincorporation events (and, optionally, actual nucleotide incorporation events) to determine which nucleotide is incorporated at a particular site/position of the template nucleic acid. The invention further provides methods that modulate (e.g., increase) the branching rate of a polymerization reaction to facilitate identifying which nucleotide is incorporated at a particular site. A nucleic acid sequencing system that detects and utilizes signal pulses or signatures from branch fraction nonincorporation events to determine the sequence of a template nucleic acid is also provided by the invention. The invention further provides compositions that include modified recombinant polymerases that exhibit properties, e.g., increased branching fraction, delayed translocation or increased nucleotide or nucleotide analog residence time, which can be particularly desirable for these applications. These improved polymerase properties can facilitate readout accuracy. In addition, the invention provides methods of generating the modified polymerases of the invention and methods in which such polymerases can be used to e.g., sequence a DNA template and/or make a DNA.

[0009] In one aspect, the invention provides methods for determining which labeled nucleotide is incorporated at a particular site during a template dependent polymerization reaction. The methods include incorporating the nucleotide into a nucleic acid polymer, whereby signal pulses or signatures are generated from branch fraction nonincorporation events and, optionally, actual nucleotide incorporation events for the site. The methods additionally include monitoring a time course of at least branch fraction signal pulses or signatures produced by the polymerization reaction and assigning which nucleotide is incorporated at the site, using at least signal pulses or signatures from branch fraction nonincorporation sampling

events. The methods optionally comprise counting or estimating the number of redundant iterative sampling signal pulses per incorporation event, or determining an average number of redundant signal pulses per incorporation event. Optionally, the polymerization reaction is a high branch fraction polymerization reaction, where the branch fraction is optionally 70% or more, 80% or more, or 90% or more.

[0010] The methods described above optionally include at least one species of metal ion, which metal ion increases the frequency of branch fraction nonincorporation events in the reaction. Example metal ions include: Mg^{++}, Mn^{++}, Zn^{++}, Co^{++}, Ca^{++}, Fe^{++}, Cr^{++} and Sr^{++}. The methods described above optionally comprise both Mg^{++} and Mn^{++}, e.g., where the concentration of Mg^{++} is higher than the concentration of Mn^{++}.

[0011] The methods described above optionally include a Φ29, B103, GA-1, PZA, Φ15, BS32, M2Y, Nf, G1, Cp-1, PRD1, PZE, SF5, Cp-5, Cp-7, PR4, PR5, PR722, L17, T4 or T7 DNA polymerase, or a modified recombinant DNA polymerase thereof. The modified recombinant polymerase can optionally exhibit a higher branching fraction as compared to a corresponding wild-type polymerase, or an increased exonuclease rate that is about 10% to 50% as compared to its polymerization rate.

[0012] The methods described above optionally include branch fraction nonincorporation events that comprise iterative sampling of labeled unincorporatable nucleotide analogs, optionally including actual nucleotide incorporation events that comprise incorporation of unlabeled nucleotides. Actual nucleotide incorporation events optionally include incorporation of nucleotides that are differentially labeled as compared to the unincorporatable nucleotides. In one example, labeled unincorporatable nucleotide analogs comprise a link between an alpha and beta phosphate group that is not hydrolyzable by a polymerase enzyme.

[0013] The methods described above optionally include a polymerase enzyme, polymerase reaction conditions, and/or polymerase reaction substrates that are selected such that the polymerization reaction exhibits two kinetically observable steps within an observable phase of the polymerase reaction. The two kinetically observable steps are optionally steps which proceed in a bright phase or a dark phase, and the polymerase enzyme optionally

comprises a modified recombinant Φ29, B103, GA-1, PZA, Φ15, BS32, M2Y, Nf, G1, Cp-1, PRD1, PZE, SF5, Cp-5, Cp-7, PR4, PR5, PR722, L17, T4 or T7 DNA polymerase. Optionally, the polymerase reaction conditions can include, e.g., a selected metal cofactor concentration, a selected pH, a selected temperature, an enzyme activity modulator, D_2O, an organic solvent, and a buffer.

[0014] The methods described above optionally comprise branch fraction nonincorporation events that comprise noncognate branch fraction nonincorporation events. Branch fraction signal pulses or signatures are optionally generated from noncognate branching events of a nucleotide analog, e.g., guanine and thymine.

[0015] The reaction of the methods described above is optionally reacted in a DNA sequencing system, where the DNA sequencing system optionally comprises a zero mode waveguide or nanohole.

[0016] Assigning the nucleotide in the methods described above optionally comprises applying a statistical model to the signal pulses or signatures generated from branch fraction nonincorporation events, signal pulses generated from actual nucleotide incorporation events, or both, which statistical model assigns a likelihood that the signal pulses or signatures correspond to an incorporation event.

[0017] The methods described above optionally comprise performing an additional template dependent polymerization reaction under high processivity reaction conditions, monitoring a second time course of signal pulses or signatures produced by the additional polymerization reaction, and compiling sequencing information derived from the second time course of signal pulses or signatures with sequencing information derived from the time course of branch fraction signal pulses or signatures.

[0018] In another aspect, the invention provides multi-modal sequencing methods that comprise performing a first template dependent sequencing reaction in a first mode comprising a first set of reaction conditions and collecting initial sequencing information produced by the first sequencing reaction. Additionally, the methods can include performing a second sequencing reaction of the template, or a copy thereof, in a second mode that includes a second

set of reaction conditions and collecting additional sequencing information produced by the second sequencing reaction. The methods can include compiling the initial and additional sequencing information to provide a sequence of at least a portion of the template.

[0019] The second sequencing reaction of the methods described above is optionally produced by altering one or more reaction conditions of the first sequencing reaction, and the initial and additional sequencing information are collected in real time. Optionally, altering one or more reaction conditions comprises adding one or more polymerase cofactors to the first sequencing reaction, where the cofactors of the first sequencing reaction are optionally Mn^{++} or Mg^{++}. Both the first and second sequencing reactions comprise single template molecule sequencing reactions. For example, the first mode can produce a higher branch fraction than the second mode. The second mode optionally produces longer read lengths than the first mode. The template of the methods described above is optionally adapted to sequencing, e.g., a single circular template molecule, e.g., where the method includes switching between the first and second modes.

[0020] The invention also provides methods for determining which of two or more labeled nucleotides is incorporated at a site during a template-dependent polymerization reaction. The methods include incorporating the nucleotide into a nucleic acid polymer produced by the polymerization reaction, whereby signal pulses or signatures are generated. The methods further include monitoring the pulses or signatures, and using the presence of multiple pulses corresponding to the nucleotide, or identical molecules thereof, to assign which labeled nucleotide is incorporated at the site. The multiple pulses optionally include 2 to 20 pulses and can be generated from branch fraction nonincorporation events, which events are optionally induced by sequencing compositions that include at least one species of metal ion. Metal ions of the method can include Mg^{++}, Mn^{++}, Zn^{++}, Co^{++}, Ca^{++}, Fe^{++}, Cr^{++} and Sr^{++}. Optionally, using the presence of multiple pulses comprises distinguishing incorporation and nonincorporation signals to assign which labeled nucleotide was incorporated at the site.

[0021] In another aspect, the invention provides nucleic acid sequencing systems that, during operation of the system, sequences a nucleic acid. The nucleic acid sequencing system comprises a signal detector that detects at least signal pulses or signatures from branch fraction nonincorporation events during sequencing of a template nucleic acid, system instructions or

software that assigns a sequence based upon detection of at least signal pulses or signatures from branch fraction nonincorporation events, and a user output module that displays the sequence to the user.

[0022] Signal pulses or signatures from branch fraction nonincorporation events of the nucleic acid sequencing systems described above are optionally generated during a first sequencing mode, where the signal detector subsequently detects signal pulses generated during a low branch fraction second sequencing mode, and where the system instructions assign a sequence based upon detection of signal pulses or signatures from the first and second sequencing modes.

[0023] The nucleic acid sequencing systems optionally comprise a zero-mode waveguide or nanohole proximal to the signal detector, where during operation of the system, a sequencing reaction is contained by the zero-mode waveguide or nanohole.

[0024] In another aspect, the invention provides compositions that include a modified recombinant nucleic acid polymerase that exhibits an altered property selected from an increased branching fraction during a polymerization by the polymerase, an altered translocation property of the polymerase during a polymerization reaction, and a combination of these two altered properties, where the altered property or properties is altered as compared to a corresponding wild-type polymerase.

[0025] The modified recombinant polymerase of the compositions described above can optionally be a modified recombinant Φ29, B103, GA-1, PZA, Φ15, BS32, M2Y, Nf, G1, Cp-1, PRD1, PZE, SF5, Cp-5, Cp-7, PR4, PR5, PR722, L17, T4, or T7 polymerase. Other available polymerases can also be used as starting points for modification to alter translocation rates or to modulate branch fraction activity, such as reverse transcriptases and DNA-dependent RNA polymerases.

[0026] The modified recombinant polymerase exhibiting an increased branching fraction can optionally comprise at least one amino acid substitution or deletion or combination of substitutions or deletions selected from: N62D and Y454A; D362S; Y259H; F237Y; L381I; Y369H; H461Y; A377G; K138Q; H461D; A377S; N62D and K371Q; V118L; and K124R; where numbering of the residues is relative to a wild-type Φ29 polymerase of SEQ ID NO: 3.

The modified recombinant polymerase exhibiting an increased branching fraction can optionally exhibit increased exonuclease activity, where the polymerase exhibits an exonuclease rate that is about 10% to 50% as compared to its polymerization rate.

[0027] Modified recombinant polymerases of the compositions described above optionally exhibit a branching fraction that is at least 50% greater, at least 100% greater, or at least 200% greater than the branching fraction of a wild-type Φ29 polymerase of SEQ ID NO: 3. Optionally, the polymerases exhibit increased exonuclease activity as compared to the corresponding wild type polymerase, where the increased exonuclease activity is optionally about 10% to 50% as compared to its polymerization rate.

[0028] The modified recombinant polymerase exhibiting an altered translocation property can optionally comprise a fusion protein that comprises at least a subsequence of the parental polymerase (e.g., a Φ29 DNA polymerase) and at least one heterologous polypeptide sequence (see, e.g., SEQ ID No. 1 and SEQ ID No. 2 in the sequence listing herein). Optionally, the fusion of the at least a subsequence of the parental polymerase and the heterologous polypeptide sequence can occur at or near the c-terminal end of the parental polymerase. The wild-type polymerase is optionally a Φ29 polymerase. The heterologous polypeptide sequence can optionally comprise at least one charged amino acid, where the at least one charged amino acid can optionally be histidine or a chain of histidines. Optionally, the fusion proteins described above can comprise a linker between the at least a subsequence of the parental polymerase and the heterologous polypeptide sequence, where the linker optionally comprises a Ser3Gly linker.

[0029] The modified recombinant polymerase exhibiting an altered translocation property can optionally comprise at least one amino acid substitution or deletion or combination of substitutions or deletions selected from Asp570Lys; Asp570Ala; Asn313Lys; Asn313Ala; Gln303Lys; Gln303Ala; Gly532Ser; Met533delet; Cys530delet; Met533delet and Cys530delet; Gly532delet; Ala531Gly; Thr573Lys; Thr573Ala; Asn396Lys; Thr571Lys; Thr571Ala; Thr534Lys; Thr534Ala; Asp535Lys; Asp535Ala; Lys529Ala; and Lys529Asn; where numbering of the residue positions is relative to a wild-type Φ29 polymerase of SEQ ID NO: 3.

[0030] The altered translocation property of the modified recombinant DNA polymerases can optionally comprise a delay in translocation. Modified recombinant polymerases of the compositions described above can optionally exhibit a delay in translocation that is at least about 2.5x, 10x or 15x greater than a corresponding wild-type polymerase.

[0031] The modified recombinant polymerases of the compositions described above optionally exhibit an increased nucleotide or nucleotide analog residence time or increased processivity as compared to a corresponding wild-type polymerase.

[0032] The compositions comprising a modified recombinant polymerase that exhibits an altered property described above can include a phosphate-labeled nucleotide analog, a DNA template, and a modified recombinant DNA polymerase, e.g., any of the polymerases described above, that can incorporate the nucleotide analog into a copy nucleic acid in response to the DNA template. These compositions can be present in a DNA sequencing system, e.g., a zero-mode waveguide or nanohole. Optionally, the polymerase of the compositions can be immobilized on a surface.

[0033] In a related aspect, the invention provides methods of sequencing a nucleic acid template. The methods include providing a reaction mixture that includes the nucleic acid template, a replication initiating moiety that complexes with or is integral to the template, the modified recombinant nucleic acid polymerase of the compositions described above, where the polymerase is capable of replicating at least a portion of the template using the moiety in a template-dependent polymerization reaction, and one or more nucleotides and/or nucleotide analogs. In addition, the methods subject the reaction mixture to a polymerization reaction in which the modified recombinant polymerase replicates at least a portion of the template in a template-dependent manner, where one or more nucleotides and/or nucleotide analogs are incorporated into the resulting copy nucleic acid. The methods additionally identify a time sequence of incorporation of the one or more nucleotides and/or nucleotide analogs into the resulting copy nucleic acid. Optionally, the methods include a modified recombinant polymerase that exhibits increased processivity relative to the wild-type polymerase. The methods optionally include identifying the time sequence of incorporation by observing more than one signal pulse per nucleotide incorporation event. Subjecting the reaction mixture to a

polymerization reaction and identifying a time of sequence incorporation can optionally be performed in a zero mode waveguide, nanohole or other micro- or nano-structure.

[0034] The invention also provides methods of making a nucleic acid that include providing a reaction mixture that comprises a template, a replication initiating moiety that complexes with or is integral to the template, a modified recombinant DNA polymerase with an altered property or combination of altered properties, e.g., such as those described above, which can replicate at least a portion of the template using the moiety in a template-dependent polymerase reaction, and one or more nucleotides and/or nucleotide analogs. In addition, the methods include reacting the mixture such that the polymerase replicates at least a portion of the template in a template-dependent manner, whereby the one or more nucleotides and/or nucleotide analogs are incorporated into the resulting nucleic acid. Optionally, the methods include detecting incorporation of at least one of the nucleotides and/or nucleotide analogs, which optionally includes observing more than one signal pulse per nucleotide incorporation event. The mixture is optionally reacted in a zero mode waveguide or nanohole, and the modified recombinant polymerase optionally exhibits an increased nucleotide or nucleotide analog residence time and/or processivity as compared to the parental polymerase.

[0035] In a related aspect, the invention provides methods of making a modified recombinant DNA polymerase that include mutating a polymerase of interest, e.g., a Φ29-type DNA polymerase, and selecting resulting modified polymerases for a property selected from increased branching fraction and altered translocation. Mutating the polymerase of interest can optionally comprise structurally modeling the polymerase to identify a feature that may affect branch fraction or altered translocation. Optionally, mutating the polymerase of interest includes making a library of modified recombinant polymerases, and selecting the modified polymerases includes screening the library to identify at least one member exhibiting the property. The polymerase of interest optionally includes a Φ29, B103, GA-1, PZA, Φ15, BS32, M2Y, Nf, G1, Cp-1, PRD1, PZE, SF5, Cp-5, Cp-7, PR4, PR5, PR722, L17, T4 or T7 polymerase. Modified recombinant polymerases that exhibit the property optionally exhibit increased nucleotide or nucleotide analog residence time, or increased processivity, as compared to a corresponding wild-type polymerase.

BRIEF DESCRIPTION OF THE DRAWINGS

Figure 1 schematically illustrates a system of the invention.

Figure 2 schematically illustrates a sequencing by incorporation reaction and the resulting characteristics of signal pulses detected by a system that employs reaction conditions or polymerases of the invention.

Figure 3 is a time sequence of signal pulses generated from a sequencing by incorporation reaction under relatively high branch fraction conditions.

Figure 4 is a schematic illustration of a sequencing by incorporation reaction in which unincorporatable nucleotides are included in the reaction.

Figure 5 shows a theoretical representation of the frequency of binding events per incorporation for a polymerase reaction having one rate-limiting step or two rate-limiting steps within an observable phase.

Figure 6 is a schematic illustration of the reaction cycle for polymerase-mediated nucleic acid primer extension.

Figure 7 schematically illustrates a system of the invention that utilizes more than one mode of sequencing.

DETAILED DESCRIPTION

[0036] The invention is generally directed to modified or engineered compositions that are characterized by modified profiles or characteristics for incorporation of nucleotides in template directed nucleic acid synthesis. Such characteristics include, for example, increased frequency of branching events, changes in reaction rates that lead, e.g., to delayed polymerase translocation and/or increased nucleotide or nucleotide analog retention time during polymerization reactions. Individually or in combination, these modifications can increase sequence readout accuracy (e.g., increase sequence accuracy in single molecule sequencing reactions) using the methods of the invention. Polymerases of the invention optionally also include additional mutations or modifications that provide other desirable features, e.g., modify one or more kinetic features of the polymerase (e.g., increased processivity), increased surface stability for polymerases bound to a surface, or the like.

[0037] During sequencing by incorporation, e.g., single molecule sequencing by synthesis (SMS), nucleotide (or nucleotide analog) incorporation events are detected in real-time as the bases are incorporated into the extension product. This can be accomplished by immobilizing a synthesis complex, which includes a polymerase enzyme, such as a DNA polymerase enzyme, a template nucleic acid sequence, and a primer sequence that is complementary to a portion of the template sequence, within an optically confined space or otherwise resolved as an individual molecular complex. Some SMS methods employ nucleotide analogs that include fluorescent labels coupled to the polyphosphate chain of the analog, which are then exposed to the complex. Upon incorporation, the nucleotide – along with its fluorescent label – is retained by the complex for a time and in a manner that permits the detection by a sequencing system of a signal "pulse" from the fluorescent label at the incorporation site. The sequentially detected signal pulses are then interpreted by the sequencing system to generate a readout corresponding to the sequence of the template nucleic acid. For a discussion of preferred sequence by incorporation processes, see, e.g., U.S. Patent Nos. 6,056,661, 7,052,847, 7,033,764, 7,056,676, 7,361,466, the full disclosures of which are hereby incorporated herein by reference in their entirety for all purposes. Further details regarding base calling during sequencing by incorporation methods are found in Tomaney et al. PCT Application Serial No. PCT/US2008/065996 METHODS AND PROCESSES FOR CALLING BASES IN SEQUENCING BY INCORPORATION METHODS, incorporated herein by reference in its entirety for all purposes.

[0038] **Figure 2** schematically illustrates a sequencing by incorporation reaction and the resulting patterns of signal pulses detected by a system that employs reaction conditions or polymerases of the invention. Figure 2A schematically illustrates a polymerization reaction where dye-labeled nucleotides are incorporated in a stepwise fashion according to the sequence of the template strand. When a dye-labeled nucleotide enters the detection region (dashed box) which encompasses the polymerase, the dye emits optical signal pulses or signatures in response to excitation radiation that are detected by a signal detector. Examples of detection methods and optically confined reaction regions include, e.g., Zero Mode Waveguides, e.g., as described in U.S. Patent Nos. 6,917,726, 7,013,054, 7,181,122, and 7,292,742, the full disclosures of which are hereby incorporated by reference in their entirety for all purposes.

Figure 2B schematically illustrates the patterns or characteristics of signal pulses that would arise under standard conditions (Panel I), increased branching conditions (Panel II), conditions that include a polymerase with a decreased translocation rate (Panel III) and conditions that include a polymerase that exhibits increased nucleotide analog residence time (Panel IV). The resulting patterns or characteristics of signal pulses from the various conditions are described in detail below.

I. INCREASED BRANCHING

[0039] "Branching" is a phenomenon that occurs during polymerization. During a polymerase kinetic cycle, sampling of each of four possible nucleotides (or nucleotide analogs) occurs until a correct Watson-Crick pairing is generated (see, e.g., Hanzel et al. WO 2007/076057 POLYMERASES FOR NUCLEOTIDE ANALOG INCORPORATION for an example model description of the kinetic cycle of a polymerase). However, chemical linkages between a sampled nucleotide and a 3'OH group of a preceding base can fail to occur for a correctly paired nucleotide, due, e.g., to release of the correctly paired base from the active site. This can occur as a result of the nucleotide leaving the site without a covalent bond being formed, or e.g., as a result of cleavage of the covalent bond (e.g., due to exonuclease activity) prior to polymerase translocation to the next incorporation site. During single molecule sequencing (SMS) procedures, and particularly those single molecule processes that monitor incorporation in real time, where both the failed incorporation and the actual incorporation of the nucleotides provide signal pulses, sequences deciphered for the incorporation site can have an incorrect "insertion" relative to the correct sequence as a result of such branching. This phenomenon is termed "branching" because it leads to a branch in the sequence (a site where two identical molecules will be read as having different sequences) and may lead to increased error rates during SMS.

[0040] While branching can, in many applications of single molecule sequencing processes, be viewed as an accuracy reducing phenomenon, in at least a first aspect of the present invention, increased branching is exploited to increase sequence accuracy by providing redundant signal events resulting from iterative sampling of labeled nucleotides or

nucleotide analogs. In particular, improved sequence reliability and accuracy is achieved by providing reaction conditions and/or polymerases that exhibit a relatively high average branching fraction for a particular nucleotide or nucleotide analog and a certain distribution of branch signal pulses around this average. Such compositions are used in combination with a sequencing system that observes and interprets more than one signal pulse or signature per incorporation event to identify the nucleotide sequence of a target or template nucleic acid. This is advantageous in the present invention because detecting more than one signal pulse or signature per incorporation event provides inherent redundancy of signal for each desired incorporation event. In some cases, the "signature" will include regions of optical signal versus time that are characteristic of the branching nucleotide, but do not appear as individual pulses. This can occur, for example, when a sequence of pulses are not individually resolved. Further details regarding sequencing under high branch fraction conditions can be found in Bjornson et al. PCT Application Serial No. PCT/US2009/000921 COMPOSITIONS AND METHODS FOR USE IN ANALYTICAL REACTIONS, incorporated herein by reference in its entirety for all purposes. Additional information useful to sequencing under high branch fraction conditions can be found in Bjornson et al. PCT Application Serial Number PCT/US2009/002003 TWO SLOW-STEP POLYMERASE ENZYME SYSTEMS AND METHODS, incorporated herein by reference in its entirety for all purposes.

[0041] The branching fraction is the proportion of cognate nucleotide (or nucleotide analog, e.g., A488dA4P) dissociation events from the polymerase active site as compared to the total number of events, e.g., the sum of the dissociation events and the incorporation events for the cognate nucleotide or nucleotide analog. The present invention provides high branch fraction polymerization reactions. As used herein, a high branch fraction polymerization reaction includes a reaction that exhibits a branching fraction of at least about 70% or more, about 80% or more, about 85% or more, about 90% or more, or about 95% or more. For example, in a sequencing reaction in which the branching fraction is 80%, 80% of the total interactions of the nucleotide or nucleotide analog with the polymerase binding pocket result in dissociation, rather than incorporation, of the nucleotide or nucleotide analog.

[0042] An aspect of the invention is a method of nucleic acid sequencing by monitoring an optical signal from a polymerase reaction, wherein the base call, or the assignment of the incorporated base is made on the basis of multiple pulses from the same nucleotide. The number of pulses used to assign which base has been incorporated may depend on the branching fraction under the conditions of the polymerase reaction. In some cases, the number of pulses used to assign which nucleotide has been incorporated will vary between the different nucleotides in that reaction medium. The number of pulses used to assign which nucleotide is incorporated can be about 2, 3, 4, 5, 6, 7, 8, 9, 10, 12, 15, 20, 30, or more pulses. The number of pulses can be between about 2 and about 30 pulses, between about 2 and about 20 pulses, or between about 3 and about 25 pulses. The number of pulses can be expressed, for example, as the average number of pulses used to assign a given nucleotide for one sequencing reaction.

[0043] Sequence read errors during SMS can also derive from the incorporation of nucleotides or nucleotide analogs that constitute dark matter (for the purposes of this disclosure, "dark matter" refers to unlabeled nucleotides or nucleotide analogs with nonfunctional labels, e.g., broken fluorophores). Here, a genuine incorporation event is not detected due to the absence of a signal pulse from the dark matter, and a subsequent incorporation event is interpreted by the sequencing system as occurring at the position where the dark matter was incorporated. Dark matter, therefore, may potentially contribute to error rates in single molecule sequencing that utilizes the incorporation of labeled nucleotides.

[0044] In certain aspects, the reaction conditions, modified recombinant polymerases, and/or nucleotide analogs of the present invention – employed in conjunction with the sequencing system of the present invention – reduce sequence read errors that might result from missed pulses. The reaction conditions induce a relatively high branch fraction polymerization reaction – and the modified recombinant polymerases exhibit increased average branching fractions - such that a greater number of nucleotide analogs, which, if incorporated, would correctly pair with the corresponding nucleotide of the template strand, enter the active site before an analog is eventually incorporated into the extension product. The nucleotide analogs that enter the active site, but fail to incorporate, produce redundant signal pulses or signatures at each incorporation site, resulting in multiple redundant signal

events for each incorporation event. An example of signal pulses generated under relatively high branch fraction conditions is shown in **Figure 3**.

[0045] The sequencing system takes into account the average branching fraction of the polymerase and a certain distribution of branch pulses or signatures per nucleotide incorporation around this average. Because multiple signal pulses are observed for each incorporation event, branching events involving unlabeled nucleotides or nucleotides with nonfunctional labels, i.e., dark matter, do not result in a sequencing read error, but rather only slightly decrease the distribution of the average number of pulses or signatures per incorporation. In the event that dark matter is incorporated into the extension product, signal pulses derived from branching events involving nucleotide analogs with functional labels prior to dark matter incorporation can provide sufficient redundancy for determining the correct base at the incorporation site.

[0046] As will be appreciated, high branch fraction sequencing conditions can also be used for sequencing RNA templates, for example using reverse transcriptase enzymes and for RNA synthesis, for example by DNA dependent RNA polymerases.

A. Enhanced Sequencing Using Reaction Conditions That Promote Branching

[0047] The present invention provides reaction conditions – such as the type, level, and relative amounts of cofactors – that increase the frequency of branching events during nucleic acid polymerization reactions. Such reaction conditions may be used in combination with polymerases that are engineered to exhibit increased branching fractions under selected conditions, or can be used with polymerases that are unaltered with respect to branching properties. The phosphoryl transfer reaction of DNA polymerases is typically catalyzed by a two-metal ion mechanism, where two divalent metal ions, e.g., Mg^{++} and/or Mn^{++}, complexed with the DNA polymerase facilitate the incorporation of a nucleotide into the 3'OH of the extension product. One of the metal ions is proposed to interact with the 3'OH of the primer strand, thereby facilitating its attack on the α-phosphate of the incoming nucleotide. Both metal ions are believed to stabilize the transition state that occurs during the course of the extension reaction.

[0048] During the course of the polymerase reaction, divalent metal cofactors, such as magnesium or manganese, will interact with the enzyme-substrate complex, playing a structural role in the definition of the active site. For a discussion of metal cofactor interaction in polymerase reactions, see, e.g., Arndt, et al., Biochemistry (2001) 40:5368-5375. For example, and without being bound to any particular theory of operation, it is understood that metal cofactor binding in and around the active site serves to stabilize binding of incoming nucleotides. For further details regarding the effect of metal cofactors on polymerase kinetics and nucleic acid synthesis reactions, see Bjornson et al. PCT Application Serial Number PCT/US2009/002003 TWO SLOW-STEP POLYMERASE ENZYME SYSTEMS AND METHODS, incorporated herein by reference in its entirety for all purposes.

[0049] In the context of the present invention, it has been discovered that modulation of the concentration of a divalent metal cofactor, or competitive modulation of two or more divalent metal cofactors, to the synthesis reaction can result in increased branching for enhanced nucleic acid sequencing without a consequent increase in negative reaction events. As described in detail herein, the increased branching provides redundant signal pulses or signatures, thereby reducing or eliminating the occurrence of missed signal pulses and improving sequence accuracy. As used herein, a signature can include regions of optical signal versus time that is characteristic of the branching nucleotide, but does not appear as an individual signal pulse.

[0050] In the synthesis reaction, certain divalent or trivalent metal cofactors, such as magnesium and manganese are known to interact with the polymerase to modulate the progress of the reaction (See, e.g., U.S. Patent No. 5,409,811). As will be appreciated, depending upon the nature of the polymerization reaction, environmental conditions (e.g., temperature, pH, etc.), the polymerase used, the nucleotides employed, etc., different metal co-factors will have widely varying catalytic effects upon the polymerization reaction. In the context of the present invention, different metal cofactors will be referred to herein based upon their relative catalytic impact on the polymerization reaction, as compared to a different metal included under the same reaction conditions. For purposes of discussion, a first metal cofactor that interacts with the polymerase complex to support the polymerization reaction to a higher level than a second metal cofactor under the same conditions is termed a "catalytic metal ion" or "catalytic metal."

[0051] The present invention provides sequencing compositions and methods that include, e.g., divalent metal ions at concentrations that induce high branch fraction polymerization reactions. Divalent metal ions of the invention can be, e.g., Mg^{++}, Mn^{++}, Zn^{++}, Co^{++}, Ca^{++}, Fe^{++}, Cr^{++}, and/or Sr^{++}. For the purposes of this disclosure, a high branch fraction polymerization reaction includes a reaction that exhibits a branching fraction of about 70% or more, about 80% or more, about 85% or more, about 90% or more, or about 95% or more.

[0052] In one embodiment, a high branch fraction polymerization reaction is induced in a sequencing composition that includes a DNA polymerase, e.g., a Φ29-derived polymerase that uses Mn^{++} as the sole source of metal cofactor at a concentration of about 250 μM or less, about 200 μM or less, about 150 μM or less, about 125 μM or less, about 100 μM or less, about 75 μM or less, or about 50 μM or less. By contrast, branching is typically not promoted in sequencing compositions that include, e.g., about 500 μM Mn^{++} or more, in the absence of other factors.

[0053] In another embodiment, a high branch fraction polymerization reaction is induced by a sequencing composition that includes a DNA polymerase, e.g., a Φ29-derived polymerase that uses Mg^{++} as the sole of metal cofactor at a concentration of about 1 mM or more, about 2 mM or more, about 3 mM or more, about 5 mM or more, about 10 mM or more, about 20 mM or more, about 30 mM or more, about 40 mM or more, or about 50 mM or more.

[0054] In yet another embodiment, a high branch fraction polymerization reaction is induced by a sequencing composition that includes a DNA polymerase, e.g., a Φ29-derived polymerase that uses both Mg^{++} and Mn^{++} as metal cofactors, where both the absolute values and ratio of the two concentrations of Mg^{++} and Mn^{++} determines the extent of branching. For example, the present invention provides sequencing compositions that induce branching by including both Mg^{++} and Mn^{++}, where the concentration of Mg^{++} is greater than the concentration of Mn^{++}. In one particular embodiment, the sequencing composition includes about 10 mM $MgCl_2$ and about 100 μM $MnCl_2$. A range of suitable concentrations to increase branching are 0-200 mM $MgCl_2$ and 0.01-50 mM $MnCl_2$, and all possible combinations of values between those two ranges.

[0055] In another embodiment, the reaction conditions include Mn^{++} and a metal cofactor other than Mg^{++}, such as calcium, barium, strontium, iron, cobalt, nickel, tin, zinc, and europium. For example, these metals can be added to the polymerization reaction in salt form such as $Sr(OAc)_2$, $CoCl_2$, $SnCl_2$, $CaCl_2$, or $ZnSO_4$. Both the absolute values and ratio of the two concentrations can influence the extent of branching.

[0056] The present invention also provides methods for increasing the branching fraction during sequencing by incorporation by modifying reaction conditions other than the identity and/or concentrations of metal ions. For example, the pH (lowering the pH to about 6.5), temperature (e.g., decreased temperature), addition of D_2O, and/or addition of small molecule inhibitors (e.g., a noncompetitive inhibitor that slows covalent attachment of the nucleotide to the 3-OH of the growing strand of the nucleic acid being synthesized, e.g., a non-competitive HIV-RT inhibitor), can be used to alter the branching fraction of the polymerization reaction.

[0057] As will be appreciated, the sequencing compositions and methods described above that utilize metal cofactors to induce relatively high branching fractions can be used in combination with any other embodiments described herein, including: (1) modified recombinant polymerases that exhibit increased branching fractions as compared to the corresponding wild-type polymerases; (2) iterative sampling of unincorporatable nucleotides; (3) two slow-step enzyme systems; (4) detection of noncognate branching events; (5) modified recombinant polymerases that exhibit altered translocation properties as compared to the corresponding wild-type polymerases; (6) modified recombinant polymerases that exhibit increased nucleotide residence time; (7) sequencing of nucleic acid templates using more than one mode; and/or (8) any other combination of embodiments described herein. As will also be appreciated, high branch fraction sequencing conditions can be used for sequencing RNA templates, for example using reverse transcriptase enzymes, and for RNA synthesis, for example by DNA dependent RNA polymerases.

B. Enhanced Sequencing Using Modified Recombinant Polymerases with Increased Branching Fractions

[0058] During a polymerase kinetic cycle, sampling of each of the possible nucleotides or nucleotide analogs occurs until a correct Watson-Crick pairing is generated. According to structural studies of DNA polymerases complexed with DNA substrates, the primer-terminus does not typically form a covalent bond with an incorrectly paired nucleotide (Berman, et al. (2007) "Structures of phi29 polymerase complexed with substrate: the mechanism of translocation in polymerases." *EMBO J* **26:** 3494–3505). Conversely, branching events can occur during the polymerase kinetic cycle, where chemical linkages between a correctly paired nucleotide and a 3'OH of a preceding base fail to form, e.g., due to premature release of the sampled nucleotide from the active site. The kinetic cycle is then repeated for the same site, eventually resulting in the physical incorporation of the correct nucleotide.

[0059] As described above, these branching events can result in sequence read errors in standard sequencing methods, e.g., due to extra incorporation signals relative to the template sequence, received by a sequencing system that monitors signal pulses from the nucleotide analog at the active site as a proxy for incorporation, if the system does not account for the branching events. However, a sequencing system that utilizes branching events and calls bases according to multiple signal pulses or signatures for each incorporation event can be used in combination with polymerases that exhibit a high average branch fraction to improve sequence read accuracy. Under such conditions, redundant signals pulses or signatures resulting from iterative sampling of a labeled cognate nucleotide or nucleotide analog can reduce the error rate of a sequence read, as compared to lower branch fraction sequencing conditions where only one or a small number of signal pulses for each incorporation event can go undetected by the sequencing system.

[0060] The present invention provides modified recombinant polymerases and reaction conditions with increased branching fractions that can be used to improve sequence read accuracy. The branching fraction is the proportion of cognate nucleotide (or nucleotide analog, e.g., A488dA4P) dissociation events from the polymerase active site as compared to the total number of events, e.g., the sum of the incorporation events and dissociation events for the cognate nucleotide or nucleotide analog. Either incorporation or non-incorporation events, or both, can be detected by monitoring a signal profile produced by a sequencing reaction.

[0061] In the present invention, modification of a DNA polymerase by mutagenesis is used to increase the frequency of branching events. In exemplary embodiments, this modification may include one or more of either creating a more loosely structured binding pocket for the (typically non-natural) nucleotides that are incorporated during SMS, or by structurally modifying the polymerase to increase exonuclease activity. Random mutation strategies can also be used, e.g., in conjunction with appropriate screening steps to select libraries of mutants for increased branching (or other properties of interest). Combinations of random and site-directed mutagenesis can also be used, typically in conjunction with selection of mutant libraries for a property of interest.

[0062] As will be appreciated, polymerase enzymes of the present invention are not limited to DNA polymerases. The present invention provides modified recombinant reverse transcriptase enzymes and modified recombinant DNA-dependent RNA polymerases, which can exhibit increased branching fractions during RNA template sequencing and RNA synthesis, respectively.

[0063] One class of example mutants described in this application were designed to address branching fraction by modifying various sites in, e.g., a Φ29 polymerase, predominantly in and around the binding pocket, to create weaker polymerase-analog interactions during an extension (polymerization) reaction. A second class of example mutants described in this application were designed to increase branching by modifying various sites in, e.g., a Φ29 polymerase, predominantly in and around the exonuclease domain in order to increase the exonuclease rate of the polymerase to about 10% to 50% as compared to its polymerization rate. As noted, the "branching fraction" is the proportion of cognate nucleotide (or nucleotide analog, e.g., a dye-labeled analog) dissociation events from the polymerase active site to the total number of events, e.g., the sum of the incorporation events and dissociation events. For the purposes of this disclosure, dissociation events also include cleavage of an incorporated nucleotide as a result of exonuclease activity. These mutational features, i.e., increased branching by creating a more loosely structured binding pocket or increasing exonuclease activity, can be provided in combination.

[0064] Desirably, the branching fraction for a polymerase for a given nucleotide of interest (e.g., a labeled nucleotide analog) can be more than 50%, more preferably more than 60%, yet more preferably more than 70%, and still more preferably more than 80% or more of the total interactions, e.g., dissociation events and association events, of the nucleotide analog with the polymerase binding pocket. In comparison, a parental Φ29 polymerase exhibits a branching fraction of approximately 23% for, e.g., a thymine nucleotide analog that includes an Alexa568 fluorophore (Invitrogen Inc., Carlsbad, CA) linked to the terminal phosphate of a hexaphosphate chain, also referred to as A568dT6P, wherein approximately 23% of the total events with a gamma-linked A568dT6P nucleotide analog in the polymerase binding pocket are dissociation events.

[0065] The invention provides methods for generating recombinant polymerases that comprise modifications that increase the frequency of branching, which can be useful in any number of applications where accuracy of polymerization is beneficial, e.g., high-throughput sequencing systems, e.g., in a nanohole (an aperture of less than 1 μM diameter through which a synthesis complex can be illuminated by optical energy or monitored electrochemically) or specialized nanoholes such as zero-mode waveguides (ZMW), SNP genotyping using single base extension methods, real time monitoring of amplification, e.g., RT-PCR methods, and the like. Also provided by the invention are compositions that include such polymerases and methods in which these polymerases can be useful in, e.g., sequencing or making DNA.

[0066] In some embodiments, the compositions can also include nucleotide analogs, and preferably, optically labeled, e.g., fluorescently labeled, nucleotide analogs. In particularly preferred aspects, the compositions will include one or more types of phosphate-labeled nucleotide analog or analogs, e.g., a nucleotide analog comprising from 3-7 phosphate groups which in preferred cases may include a fluorophore coupled to the phosphate chain that is released upon incorporation, which can be incorporated into a copy nucleic acid by the modified polymerase in response to a DNA template. In some embodiments, the compositions can be present in a sequencing system, e.g. in a nanohole or specialized nanohole such as a zero-mode waveguide, where a polymerase of the invention can optionally be immobilized on a surface.

[0067] Modification of a polymerase, e.g., any of the polymerases described herein, or polymerases homologous to those described herein, by any one or more the strategies described herein can increase the frequency of branching events by creating a more loosely structured binding pocket for non-native nucleotides. The modified polymerases can comprise at least one amino acid substitution or a combination of amino acid substitutions relative to the parental polymerase.

[0068] Polymerases of the invention can be described or analyzed based upon comparison to a reference (e.g., parental) polymerase. For example, a parental polymerase (e.g., a wild type polymerase from which the polymerase of the invention is derived, or an available mutant) can serve as a reference polymerase. Comparisons between a reference polymerase and a polymerase of the invention are performed under selected reaction conditions. In general, standard reaction conditions can be defined based on the usual reaction conditions (e.g., optimized for the reaction at issue) for a given parental (e.g., wild-type) polymerase. That is, the reaction condition preferences for many polymerases are known; the appropriate adaptations of these conditions to specific applications such as SMS are known or can be determined through routine optimization or reaction conditions, and these "application optimized" reaction conditions can be used for comparison of the polymerase of the invention to the reference polymerase. For example, reaction conditions can be optimized for a reference polymerase such as a Φ29 polymerase, e.g., in an SMS application, with the reaction conditions being selected for optimal processivity, optimal fidelity, increased or decreased branch fraction, or a combination thereof, with that reference polymerase being used in a comparison to a polymerase of the invention under the optimized conditions. For example, in one SMS application, the reaction conditions can include those described above, and comparison to the polymerase of the invention can be conducted by performing a comparative assay, using the methods herein.

[0069] A number of specific examples of modified polymerases, e.g. modified to increase the average branching fraction, are described herein. The binding pocket is a portion of the polymerase that encompasses the nucleotide binding site and analog base during the pairing of a nucleotide analog with a template DNA. Because of the physical proximity of the binding pocket to the incoming nucleotide or nucleotide analog, mutations to this region can

affect the branching fraction. However, mutations that increase the branching fraction are not limited to this area of the polymerase. For example, relative to a parental or wild-type Φ29 DNA polymerase, useful modifications can include any of the following mutations: N62D and Y454A; D362S; Y259H; F237Y; L381I; Y369H; H461Y; A377G; K138Q; H461D; A377S; N62D and K371Q; V118L; and K124R. For the purposes of this disclosure, a parental polymerase includes a wild-type or available mutant/recombinant polymerase which is additionally altered to produce the desired properties of the invention, e.g., increased branching, delayed translocation or increased nucleotide analog residence time. A list of specific useful Φ29 mutants and their branching fractions (% BF) and corresponding increases in branching fraction (% increase BF) for a particular 6P nucleotide analog relative to a reference Φ29 polymerase (N62D) is provided in Table A below.

Table A

	A568dT6P		A647dG6P	
Mutation(s)	%BF	% Increase BF	%BF	% Increase BF
N62D	23	-	38	-
N62D_Y454A	30	30	48	25
D362S	31	34	46	20
Y259H	33	44	52	37
F237Y	35	53	46	21
L381I	37	60	47	22
Y369H	40	75	55	43
H461Y	43	86	57	49
A377G	44	90	75	97
K138Q	45	96	61	59
H461D	46	102	62	62
A377S	61	167	70	84
N62D_K371Q	62	172	94	145
V118L	65	181	74	92
K124R	73	216	84	119
	A555dC6P		A660dA6P	
Mutation(s)	%BF	% Increase BF	%BF	% Increase BF
N62D	27	-	22	-
N62D_Y454A	38	42	23	3
D362S	39	45	26	20
Y259H	39	44	25	11
F237Y	41	52	25	11
L381I	37	36	26	18
Y369H	43	59	28	27

	A555dC6P		A660dA6P	
Mutation(s)	%BF	% Increase BF	%BF	% Increase BF
H461Y	50	86	31	42
A377G	37	36	32	47
K138Q	48	78	27	21
H461D	51	91	31	39
A377S	72	168	56	152
N62D_K371Q	87	224	82	272
V118L	68	154	47	114
K124R	79	195	63	183

[0070] As noted, the branching fraction, e.g., % branching, is a relative measure of the number of times a correctly paired base, e.g., a Watson-Crick paired base, leaves the active site of the polymerase without forming a phosphodiester bond with the 3'OH of the primer-terminus relative to the total number of interactions that occur between a nucleotide (or nucleotide analog) and the binding pocket of the polymerase, e.g., the total number of opportunities the nucleotide or nucleotide analog has to correctly pair and incorporate. Additionally, for the purposes of this disclosure, branching refers to cleavage and dissociation from the polymerase active site of an incorporated nucleotide as a result of exonuclease activity. Branching is expressed as a percentage of the dissociation events vs. the total sum events, e.g., dissociation and association events. For example, for a polymerase harboring the Y369H mutation, for every 100 times an A568dT6P analog (i.e., a thymidine hexaphosphate nucleotide in which the terminal phosphate is labeled with an Alexa568 dye) interacts with the binding pocket of this polymerase, 40 of the events are non-productive dissociation events, e.g., wherein the analog dissociates from the polymerase instead of participating in a polymerization reaction. For this polymerase, the percent increase in branching fraction is 75% as compared to a reference phi29 polymerase (N62D) under identical reaction conditions.

[0071] The branching fraction can be measured by loading a polymerase active site with a cognate-matching nucleotide analog that can bind in the +1 and +2 positions. In the absence of divalent cation, this nucleotide cannot be incorporated into the DNA strand, so will pair with the template nucleotide at the +1 position but be released at some frequency specific for that analog/polymerase combination, e.g., the branching rate. This loading reaction is then followed by addition of a divalent cation that supports extension, e.g., Mn^{2+}, and a terminating–

type nucleotide analog, e.g., a dideoxynucleotide, comprising the same base as the cognate-matching analog in the loading step.

[0072] The dideoxy-analog will be incorporated into any +1 sites that are unoccupied and, once added, preclude further extension. Hence, polymerase active sites that are already occupied by a paired analog base extend to the +2 position, those that are not occupied (i.e. "branched") incorporate the dideoxy-type analog at +1 and do not extend, resulting in a single base addition. The extension products of this reaction are visualized by standard separation methods, e.g., gel or capillary electrophoresis, and the ratio of terminated product that is generated when a dideoxynucleotide is incorporated at the +1 position divided by the total terminated product, e.g., when a dideoxynucleotide is incorporated at both the +1 and +2 positions, indicates the fraction of 'branched' events that occur.

[0073] The branching fraction exhibited by modified polymerases of the present invention, e.g., a modified Φ29 polymerase, can be greater than a branching fraction exhibited by the corresponding wild-type polymerase for a given nucleotide analog. For example, a modified recombinant polymerase of the invention can exhibit an increased branching fraction that is greater than about 20% for the phosphate-labeled analog, greater than 50% for the phosphate-labeled analog, greater than 75% for the phosphate-labeled analog, greater than 100% for the phosphate-labeled analog, greater than 150% for the phosphate-labeled analog, or greater than 200% for the phosphate-labeled analog, as compared to the corresponding wild-type polymerase, e.g., a wild-type Φ29 polymerase, under the standard reaction conditions described above.

[0074] In some embodiments, the modified polymerase that exhibits an increased frequency of branching can also exhibit a K_m for a given phosphate-labeled nucleotide analog, e.g., any of the phosphate-labeled nucleotide analogs described herein, that is useful to achieve increased branching. For enzymes obeying simple Michaelis-Menten kinetics, kinetic parameters are readily derived from rates of catalysis measured at different substrate concentrations. The Michaelis-Menten equation, $V=V_{max}[S]([S]+K_m)^{-1}$ relates the concentration of uncombined substrate ([S], approximated by the total substrate concentration), the maximal rate (V_{max}, attained when the enzyme is saturated with substrate), and the

Michaelis constant (K_m, equal to the substrate concentration at which the reaction rate is half of its maximal value), to the reaction rate (V). To determine a K_m for a particular analog a series of extension reactions are performed with a varying concentration of the analog of interest with a fixed, saturating concentration of native nucleotides. A fit of the rate versus the substrate concentration generates estimation of the $-K_m$ as the slope of this line.

[0075] The present invention also provides polymerases with increased exonuclease activity for increasing the branching fraction during a template-dependent polymerization reaction, e.g., SMS. In a preferred aspect, a polymerases of the invention exhibits an exonuclease rate that is between about 10% and 50% as compared to its polymerization rate.

[0076] As will be appreciated that the above-identified modified or recombinant polymerases that display increased branching fractions may optionally include additional modifications that confer other useful properties described herein, e.g., delayed translocation, increased nucleotide analog residence time and/or increased processivity. As will also be appreciated, the above-identified modified or recombinant polymerases that display increased branching fractions can be used in combination with any other embodiments described herein, including: (1) reaction conditions that increase the frequency of branching fraction nonincorporation events; (2) iterative sampling of unincorporatable nucleotides; (3) two slow-step enzyme systems; (4) detection of noncognate branching events; (5) modified recombinant polymerases that exhibit altered translocation properties as compared to the corresponding wild-type polymerases; (6) modified recombinant polymerases that exhibit increased nucleotide residence time; (7) sequencing of nucleic acid templates using more than one mode; and/or (8) any other combination of embodiments described herein.

C. Enhanced Sequencing Using Iterative Sampling of Unincorporatable Nucleotide Analogs

[0077] The present invention also employs nucleotide based competitive reagent compositions for identifying sequence elements, despite not being incorporated in a nascent nucleic acid strand. In particular, the unincorporatable nucleotide analogs of the invention, while not being incorporatable, may be nonetheless capable of specifically associating with the polymerase enzyme. That is, the polymerase will sample the unincorporatable nucleotides,

retaining them within the active site for a greater length of time than nucleotides that are not complementary to the position in the template nucleic acid, and release them when they cannot be incorporated. By providing different types of nucleotide or nucleoside analogs, e.g., mimetics of A, G, T, C, and/or U, bearing distinguishable labels, e.g., spectrally resolvable fluorophores or other labeling groups, one can monitor the sampling of these nucleotides as an indication of the nucleotide that is next to be incorporated. For example, one may provide labeled, unincorporatable nucleotide analogs at concentrations in excess of incorporatable nucleotides, e.g., 2X, 5X or even 10X or greater. Each incorporation of an incorporatable nucleotide will, by virtue of the excess concentration, be preceded by repeated sampling events of the unincorporatable nucleotides, which will each carry its associated signal event. The incorporatable nucleotides may then either bear no label, or preferably, bear a label that is distinguishable from the unincorporatable nucleotides, so as to mark the termination of the sampling of a given base and proceeding onto the next base in the sequence. In such cases, it may be desirable to label all incorporatable nucleotides with a single type of fluorophore, i.e., indistinguishable from the label groups on the other types of incorporatable nucleotides present, but distinguishable from all of the unincorporatable nucleotides.

[0078] The signal detection for the foregoing process is schematically illustrated in **Figure 4**. In particular, Figure 4 shows a schematic illustration of a set of signal traces from a single molecule sequence by incorporation reaction. As shown, the plot shows five signal traces, one for each type of differentially labeled unincorporatable nucleotide analog (indicated as A', T', G' and C', as well as a trace for the signal associated with the type of label coupled to the incorporatable nucleotide (labeled as "I"). As shown, repeated sampling of the cognate unincorporatable nucleotide analog, e.g., A', provides an iterative set of signal events 402, followed by a signal 404 on the I trace indicating conclusion of the incorporation event. This pattern is repeated for the next base to be incorporated (indicated by iterative signals 406 in the T' trace, followed again by the incorporation signal 408, in the I trace, and again by the iterative sampling signal 410 in the A' trace followed by the incorporation signal 412 in the I trace. Because these unincorporatable nucleotides are mimetic of the base to be incorporated, they possess a longer retention time in the active site than the analog that is not complementary to the next base in the template, and as such, provide a signal profile that is distinguishable

from random, incorrect sampling, e.g., as indicated by transient signal events 414. Such iterative sampling may include two, three, four, five, ten or greater than ten signal events for each incorporation.

[0079] As noted above, the competitive reagents used are going to be non-reactive in the reaction of interest. In preferred aspects, and without being bound to any particular theory of operation, the competitive compounds may possess structures similar to nucleotides or portions thereof, such that they can competitively interact with the reaction of interest, e.g., through association with the polymerase active site. By way of example, such structures may comprise a polyphosphate component, e.g., a pyrophosphate, triphosphate, tetraphosphate, pentaphosphate, or longer phosphate chain, so that the compound mimics one or more of a nucleotide or the product of a polymerase mediated incorporation reaction, which is capable of competitively interacting with the polymerase, relative to the nucleotide analogs.

[0080] In certain preferred cases, additional components may be coupled to the polyphosphate component that mimic other portions of the nucleotide or nucleotide analog. By way of example, the polyphosphate component may be coupled to a cyclic and/or aromatic component that may structurally mimic the nucleoside component in its interaction with the polymerase. Such structures are generally illustrated by the following structure:

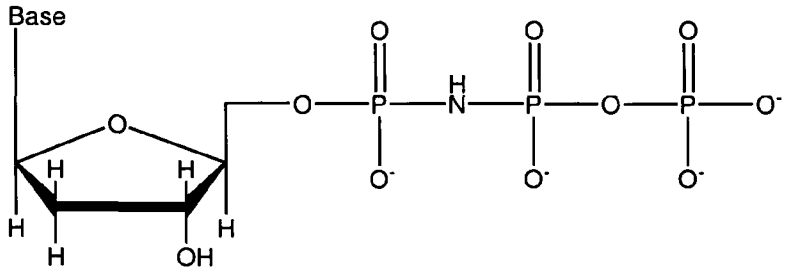
$$P\text{-}(P)_n\text{-}A;$$

where P is a phosphate or phosphonate group, n is an integer from 1 to about 6, and A includes a cycloalkyl or aryl group, a carbohydrate group, or the like.

[0081] In the case of nucleotide analogs used in analytical primer extension reactions, e.g., in nucleic acid sequence analysis, such nucleotide analogs will be unincorporatable in such primer extension reaction by the polymerase used. Further, in preferred aspects, such unincorporatable analogs will typically still be capable of interaction with the polymerase, e.g., active site binding, but will be unable to be incorporated in a primer extension reaction. In preferred aspects, this is accomplished by providing nucleotide analogs that possess nonhydrolyzable groups within the phosphate chain, such that the phosphoester linkage between the analog and the primer strand, cannot be formed, as mediated by the polymerase. One particularly effective approach to producing an unincorporatable nucleotide analog

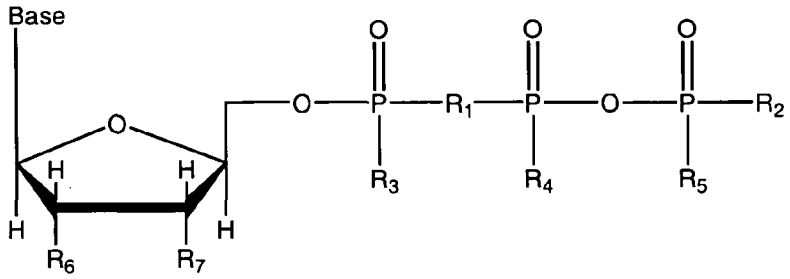
includes replacing the phosphoester linkage between the alpha and beta phosphate of a nucleoside polyphosphate with a nonhydrolyzable linkage.

[0082] One example of such an analog is illustrated below, where the oxygen group between the alpha and beta phosphate groups is replaced with an nonhydrolyzable linkage, such as the illustrated amino group.



[0083] Although illustrated as an amino linkage, it will be appreciated that a variety of other linkages may be used between the alpha and beta phosphates, e.g., an amino, methyl, thio, or other linkages not hydrolyzed by polymerase activity. Additionally, although illustrated as including three phosphate groups analogous to a nucleoside triphosphate, it will be appreciated that other polyphosphate configurations may be employed in the invention, including, for example, tetraphosphate analogs, pentaphosphate analogs, hexaphosphate analogs, and the like.

[0084] Thus, the structures employed in certain preferred aspects of the invention may generally be described with reference to the following structure:



where R_1 comprises a linking group that is non-hydrolyzable by the polymerase enzyme being used. Particularly preferred linkages include amino linkages, alkyl linkages, e.g., methyl, and thio linkages. While R_2 may comprise oxygen, in some preferred aspects, it will include additional phosphate groups, e.g., mono-, di-, or triphosphate groups coupled to the gamma

phosphate group. Alternatively or additionally, the R_2 group may include, in addition to or instead of additional phosphate groups, labeling functionalities that provide for the detection of the competitive substrates, but still permit its distinguishing from the incorporatable nucleotides. In other aspects, the R_2 group may include moieties that provide other functionalities to the reaction system other than as a labeling group. For example, R_2 may comprise an agent that reduces the potential for photodamaging effects on a polymerase enzyme, either coupled directly to the terminal phosphate group, or through a linking group.

[0085] The relative concentration of the competitive substrates to the incorporatable substrates, within a reaction mixture may generally be varied in accordance with a desired application. In particular, because the concentration of the competitive substrates affects the interactions of the complex with the incorporatable nucleotides, one can modulate those interactions by altering the ratios between incorporatable nucleotides and competitive substrates. In typical applications, however, the relative molar concentration of competitive substrate will range from about 0.5X to about 10X, 20X or greater of the concentration of the actual substrates (or incorporatable nucleotide analogs). Thus, the concentration ratio of unincorporatable nucleotide analogs to incorporatable nucleotide analogs will typically range from a lower ratio of from about 0.1:1, 0.2:1, 0.5:1 and 1:1, to an upper ratio of about 2:1, 3:1, 5:1, 10:1 or even 20:1, with each iteration of the foregoing being encompassed in the disclosure hereof.

[0086] As will be appreciated, iterative sampling of unincorporatable nucleotides can be used in combination with any other embodiments described herein, including: (1) reaction conditions that increase the frequency of branching fraction nonincorporation events; (2) modified recombinant polymerases that exhibit increased branching fractions as compared to the corresponding wild-type polymerases; (3) two slow-step enzyme systems; (4) detection of noncognate branching events; (5) modified recombinant polymerases that exhibit altered translocation properties as compared to the corresponding wild-type polymerases; (6) modified recombinant polymerases that exhibit increased nucleotide residence time; (7) sequencing of nucleic acid templates using more than one mode; and/or (8) any other combination of embodiments described herein.

D. Enhanced Sequencing Using Two Slow-Step Enzyme Systems Combined With Detection and Analysis of Branch Fraction Nonincorporation Events

[0087] Enzyme systems that exhibit kinetic mechanisms having two or more slow, kinetically observable, or partially rate-limiting reaction steps within an observable phase of the polymerase reaction can be useful for example, in single-molecule, real-time observations of such enzyme activity, which rely, at least in part, on detecting and identifying the enzyme reaction as it is occurring. By designing the reaction system to have two or more partially rate-limiting steps (i.e., "two slow-step" enzyme systems), the relative number of short, difficult to detect, events can be lowered. Details regarding enzyme systems exhibit kinetic mechanisms having two or more slow, kinetically observable, or partially rate-limiting reaction steps within an observable phase of the polymerase reaction can be found in Bjornson et al. PCT Application Serial Number PCT/US2009/002003 TWO SLOW-STEP POLYMERASE ENZYME SYSTEMS AND METHODS, incorporated herein by reference in its entirety for all purposes.

[0088] Certain types of template nucleic acid sequences present unique challenges during single molecule sequencing. For example, during single molecule sequencing of homonucleotide stretches (i.e., a portion of a template nucleic acid having two or more consecutive bases that are identical), ascertaining the number of nucleotide incorporation events that are represented by a series of signal pulses generated under high branch fraction sequencing conditions can be difficult using previous typical enzyme systems that exhibit one slow-step. By way of example, if 20 consecutive signal pulses are detected from a fluorescent-labeled nucleotide analog of dATP (e.g., A488dA4P, see below), a system that detects signal pulses and analyzes the time-sequence of those signal pulses for purposes of calling bases may not be able to accurately determine how many dATP analog incorporation events (i.e., the number of consecutive T residues in the template sequence) are represented by the 20 signal pulses. Accordingly, enzyme systems that permit a more predictable distribution of nucleotide binding events per incorporation are desirable, because, e.g., the number of nucleotides within a homonucleotide stretch can be determined by using a multiple of the expected number of binding events per incorporation.

[0089] In one aspect, the present invention provides two slow-step enzyme systems that provide a more predictable frequency or rate of nucleotide binding events (e.g., the number of times a nucleotide or nucleotide analog samples the active site of the polymerase) per incorporation event. While not being bound by any particular theory, the following theoretical basis is provided for obtaining improved single-molecule sequencing results, e.g., for homonucleotide stretches, by using a system having two or more slow steps within an observable phase. A model for the effect of two slow steps on the number of nucleotide binding events is described herein and illustrated in **Figure 5**. Figure 5 shows a plot of calculated number of binding events per incorporation for cases in which (1) one step is rate-limiting and (2) two equivalent partially rate-limiting (slow) steps are present for the observable phase in which the nucleotide is associated with the enzyme before unbinding.

[0090] For the case in which one step is rate-limiting, the distribution for the number of binding events per incorporation can be represented by the single exponential equation:

$$y = A_0 e^{-kt} \qquad \text{Eq. 1}$$

This represents the case in which, for example, one conformational change of the enzyme after nucleotide binding is the single slow step.

[0091] Figure 5 illustrates that where one slow-step is present in this phase, there is an exponentially decreasing number of binding events per incorporation, providing a distribution in which there is a relatively high probability that the number of binding events per incorporation will be low. In this scenario, it can be more challenging to distinguish homonucleotide stretches of the same base in the DNA template.

[0092] For the case in which there are two slow steps associated with binding of a nucleotide, for example via two consecutive conformational changes with similar rate constants, the number of binding events can be represented by a sum of two exponentials with an equation:

$$y = A_0 e^{-k_1 t} - B_0 e^{-k_2 t} \qquad \text{Eq. 2}$$

[0093] Figure 5 illustrates that for the case in which there are two slow steps, the probability of one or a low number of binding events per incorporation is relatively low as compared to enzyme systems having one slow step. In addition, the probability distribution for two slow steps exhibits a peak in the plot, with the most frequently observed number of binding events per incorporation greater than zero. This type of distribution can be advantageous for single-molecule sequencing where it is desired to resolve homonucleotide stretches. In this scenario, the expected number of binding events per incorporation will be a corresponding multiple factor of the most frequent occurrence of binding events per incorporation as governed by this distribution.

[0094] The two slow steps can include, e.g., nucleotide addition, enzymatic isomerization (such as to or from a closed state), and cofactor binding or release. The use of a distribution of pulses to determine a kinetic mechanism having two slow (kinetically observable) steps is described, for example, in Miyake et al. Analytical Chemistry 2008 80 (15), 6018-6022. The determination of the steps in a multistep reaction such as a polymerase reaction is described, for example, in Zhou, et al. J. Phys. Chem. B, 2007, 111, 13600-13610.

[0095] As noted above, the present invention provides enzyme systems that exhibit kinetic mechanisms having two or more slow, kinetically observable, or partially rate-limiting reaction steps within an observable phase of the polymerase reaction observable phase will generally have a time period during which the phase is observable. The time period for a bright phase, for example, can be represented by the pulse width. The time period for a dark phase can be represented, for example, by the interpulse distance. The length of each time period will not be the same for each nucleotide addition, resulting in a distribution of the length of the time periods. In some cases, the time periods with the shortest length will not be detected, leading to errors, for example in single-molecule sequencing. We have found that by designing enzyme systems such as polymerase reaction systems in which there are two slow, or kinetically observable, steps within an observable phase, the relative number of short, unobservable, time periods can be reduced, resulting in a higher proportion of observable sequencing events, and allowing for a more accurate determination of nucleotide sequence. As used herein, an observable phase includes phases that are not directly observable, but can be ascertained by measurements of other, related phases. For example, the lengths of dark phases can be

observed by measuring the times between optical pulses corresponding to a related bright optical phase. Also as described herein, a phase which is dark under some labeling conditions can be bright under other labeling conditions.

[0096] In natural polymerase-mediated nucleic acid synthesis, a complex is formed between a polymerase enzyme, a template nucleic acid sequence, and a priming sequence that serves as the point of initiation of the synthetic process. During synthesis, the polymerase samples nucleotide monomers from the reaction mix to determine their complementarity to the next base in the template sequence. When the sampled base is complementary to the next base, it is incorporated into the growing nascent strand. This process continues along the length of the template sequence to effectively duplicate that template. Although described in a simplified schematic fashion, the actual biochemical process of incorporation is relatively complex.

[0097] The process can be described as a sequence of steps, wherein each step can be characterized as having a particular forward and reverse reaction rate that can be represented by a rate constant. One representation of the incorporation biochemistry is provided in **Figure 6**. It is to be understood that the scheme shown in Figure 6 does not provide a unique representation of the process. In some cases, the process can be described using fewer steps. For example, the process is sometimes represented without inclusion of the enzyme isomerization steps 606 and 610. Alternatively, the process can be represented by including additional steps such as cofactor binding. Generally, steps which can be slow, and thus limit the rate of reaction will tend to be included. The present invention relates to methods, systems, and compositions in which the polymerization reaction has two or more slow steps within certain phases of the polymerase reaction. Various schemes can be used to represent a reaction having two slow steps that may have more or fewer identified steps. In some cases the two or more slow steps are consecutive. In some cases, there can be intervening fast steps between the two or more slow steps.

[0098] As shown in Figure 6, the synthesis process begins with the binding of the primed nucleic acid template (D) to the polymerase (P) at step 602. Nucleotide (N) binding with the complex occurs at step 604. Step 606 represents the isomerization of the polymerase from the open to closed configuration. Step 608 is the chemistry step where the nucleotide is

incorporated into the growing strand of the nucleic acid being synthesized. At step 610, polymerase isomerization occurs from the closed to the open position. The polyphosphate component that is cleaved upon incorporation is released from the complex at step 612. The polymerase then translocates on the template at step 614. As shown, the various steps can include reversible paths and may be characterized by the reaction constants shown in Figure 6 where:

[0099] k_{on}/k_{off}=DNA binding/release;

[0100] k_1/k_{-1}= nucleotide binding/release;

[0101] k_2/k_{-2}=polymerase isomerization (open/closed);

[0102] k_3/k_{-3}=nucleotide incorporation (chemistry);

[0103] k_4/k_{-4}=polymerase isomerization (closed/open);

[0104] k_5/k_{-5}=polyphosphate release/binding;

[0105] k_6/k_{-6}=polymerase translocation.

[0106] Thus, during steps 604 through 610, the nucleotide is retained within the overall complex, and during steps 604 and 606, reversal of the reaction step will yield an unproductive event, i.e., not resulting in incorporation. For example, a bound nucleotide at step 604 may be released regardless of whether it is the correct nucleotide for incorporation.

[0107] By selecting the appropriate polymerase enzyme, polymerase reaction conditions, and polymerase substrates, the absolute and relative rates of the various steps can be controlled. We have found that controlling the reaction such that the reaction exhibits two or more kinetically observable, or slow steps can produce a nucleic acid polymerization reaction in which the incorporation of the nucleotides can be observed more accurately. These characteristics are particularly useful for sequencing applications, and in particular single-molecule DNA sequencing.

[0108] In some cases, the invention involves a process having two or more kinetically observable steps that comprise steps after nucleotide binding through the step of product release. For the mechanism shown in Figure 6, this would be, for example, any of steps 606,

608, 610, and 612. In some cases, steps 608 (nucleotide incorporation) and 612 (product release) are the two slow, or kinetically observable steps. As noted previously, where one desires systems with slow steps in a dark phase, the invention may involve a process having two or more slow steps that comprise the steps after product release through nucleotide binding. For the mechanism shown in Figure 6, this would include steps 614 and 604.

[0109] In some cases, the invention involves a process in which there are two or more slow steps in two different observable phases within the polymerization, for example, two slow steps in a bright phase and two slow steps in a dark phase. For example, this could include a system having two slow steps in the steps after nucleotide binding through product release, and two slow steps for the steps after product release through nucleotide binding.

[0110] As is described herein, producing a process in which there are two or more slow steps in these portions of the polymerase reaction can result in a higher proportion of detectable enzyme states which can be useful, for example, to observe the sequential incorporation of nucleotides for nucleotide sequencing.

[0111] By the term slow-step we generally mean a kinetically observable step or partially rate-limiting step. The slow step need not be slow in the absolute sense, but will be relatively slow as compared with other steps in the enzymatic reaction. The slow, or kinetically observable steps, can be, for example, each partially rate-limiting, in that the rate of the step has a measurable effect on the kinetics of the enzymatic reaction. An enzymatic process, such as nucleic acid polymerization, can have both slower, kinetically observable steps and faster steps which can be so fast that they have no measurable effect on the kinetics, or rate, of the reaction. In some reactions, there can be a single rate-limiting step. For such reactions, the kinetics can be characterized by the rate of that single step. Other reactions will not have a single rate-limiting step, but will have two or more steps which are close enough in rate such that the characteristics of each will contribute to the kinetics of the reaction. A kinetically observable step is generally a step which is slow enough relative to the other steps in the reaction such that it can be experimentally ascertained. The experimental identification of a kinetically observable step can be done by the methods described herein, or by methods for assessing the kinetics of chemical and enzymatic reactions known in the art. For the current invention, the

slow, or kinetically observable steps, need not be the slowest step or the rate-limiting step of the reaction. For example, a process of the current invention can involve a reaction in which step 604, nucleotide addition is the slowest (rate-limiting) step, while two or more of steps 606, 608, 610, or 612 are each kinetically observable.

[0112] As used herein, the term rate, as applied to the steps of a reaction can refer to the average rate of reaction. For example, when observing a single-molecule reaction, there will generally be variations in the rates as each individual nucleotide is added to a growing nucleic acid. In such cases the rate of the reaction can be represented by observing a number of individual events, and combining the rates, for example, by obtaining an average of the rates.

[0113] As used herein, the reference to the rate of a step or rate constant for a step can refer to the forward reaction rate of the polymerase reaction. As is generally understood in the art, reaction steps can be characterized as having forward and reverse rate constants. For example, for step 608, k_3 represents the forward rate constant, and k_{-3} represents the reverse rate constant for the nucleotide incorporation. Some reaction steps, such as step 608, constitute steps which would be expected to be first order steps. Other steps, such as the forward reaction of step 604, with rate constant k_2, would be expected to be second order rate constants. For the purposes of the invention, for comparing the rate or the rate constant of a first order to a second order step, the second order rate constant k_2 can be treated as a pseudo-first order rate constant with the value [N]*k_2 where the concentration of nucleotide [N] is known.

[0114] It is generally desirable that the kinetically observable steps of the invention have rate constants that are lower than about 1000 per second. In some cases, the rate constants are lower than about 500 per second, lower than about 200 per second, lower than about 100 per second, lower than about 60 per second, lower than about 50 per second, lower than about 30 per second, lower than about 20 per second, lower than about 10 per second, lower than about 5 per second, lower than about 2 per second, or lower than about 1 per second.

[0115] In some embodiments the slowest of the two or more kinetically observable steps has a rate constant when measured under single-molecule conditions of between about 500 to about 0.1 per second, about 200 to about 0.1 per second, about 60 to about 0.5 per second, about 30 per second to about 2 per second, or about 10 to about 3 per second.

[0116] The ratio of the rate constants of each the two or more slow steps is generally greater than 1:10, in some cases the ratio of the rate constants is about 1:5, in some cases the ratio of the rate constants is about 1:2, in some cases, the ratio of rate constants is about 1:1. The ratio of the rate constants can be between about 1:10 and about 1:1, between about 1:5 and about 1:1, or between about 1:2 and about 1:1.

[0117] In some cases it is useful to consider the two slow-step system in terms of rates rather than rate constants. It is generally desirable that the kinetically observable steps of the invention have rates that are lower than about 1000 molecules per second when the reactions are carried out under single-molecule conditions. In some cases, the rates are lower than about 500 molecules per second, lower than about 200 molecules per second, lower than about 100 molecules per second, lower than about 60 molecules per second, lower than about 50 molecules per second, lower than about 30 molecules per second, lower than about 20 molecules per second, lower than about 10 molecules per second, lower than about 5 molecules per second, lower than about 2 molecules per second, or lower than about 1 molecule per second.

[0118] In some embodiments the slowest of the two or more kinetically observable steps has a rate when measured under single-molecule conditions of between about 500 to about 0.01 molecules per second, between about 200 to about 0.1 molecules per second, between about 60 to about 0.5 molecules per second, about 30 molecules per second to about 2 molecules per second, or about 10 to about 3 molecules per second.

[0119] The ratio of the rates of each the two or more slow steps is generally greater than 1:10, in some cases the ratio of the rates is about 1:5, in some cases the ratio of the rates is about 1:2, in some cases, the ratio of rates is about 1:1. The ratio can be between about 1:10 and about 1:1, between about 1:5 and about 1:1, or between about 1:2 and about 1:1.

[0120] A two or more slow-step system of the present invention can be obtained by selecting the correct set of polymerase enzyme, polymerase reaction conditions, and polymerase reaction substrates.

[0121] Table B presents exemplary Φ29 mutants that can exhibit two slow step behavior under appropriate reaction conditions. The first three modified polymerases exhibit

the most pronounced two slow step behavior, followed by the next six. As noted, the polymerases are optionally exonuclease-deficient; for example, they can also include an N62D substitution.

Table B

A484E/E375Y/K512Y/T368F
A484Y/E375Y/K512Y/T368F
N387L/E375Y/K512Y/T368F
T372Q/E375Y/K512Y/T368F
T372L/E375Y/K512Y/T368F
T372Y/K478Y/E375Y/K512Y/T368F
I370W/E375Y/K512Y/T368F
F198W/E375Y/K512Y/T368F
L381A/E375Y/K512Y/T368F
E375Y/K512Y/T368F

[0122] The polymerase reaction conditions can also be important for obtaining a two slow-step enzyme system. In particular, polymerase reaction conditions include components selected to produce two slow-step kinetics. The polymerase reaction conditions include the type and concentration of buffer, the pH of the reaction, the temperature, the type and concentration of salts, the presence of particular additives which influence the kinetics of the enzyme, and the type, concentration, and relative amounts of various cofactors, including metal cofactors. The term "polymerase reaction conditions" as used herein generally excludes the concentration of the polymerase enzyme or the concentration of the primer-template complex. Thus, two reactions are run under substantially the same polymerase reaction conditions where the first reaction has a small amount of polymerase enzyme, such as a single polymerase enzyme, and a small amount of primer template complex, such as a single primer-template complex associated with a single polymerase enzyme, and the second reaction has a higher concentration of polymerase enzyme, for example a concentration of polymerase enzyme of about 0.05 μM to 0.5 μM, and about 0.01 μM to about 0.1 μM.

[0123] It some embodiments the type and concentration of buffer are chosen in order to produce a reaction having two slow steps. Enzymatic reactions are often run in the presence of a buffer, which is used, in part, to control the pH of the reaction mixture. We have found that

in some cases the type of buffer can influence the kinetics of the polymerase reaction in a way that can lead to two slow-step kinetics. For example, in some cases, we have found that the use of TRIS as buffer is useful for obtaining a two slow-step reaction. Buffers suitable for the invention include, for example, TAPS (3-{[tris(hydroxymethyl)methyl]amino}propanesulfonic acid), Bicine (N,N-bis(2-hydroxyethyl)glycine), TRIS (tris(hydroxymethyl)methylamine), ACES (N-(2-Acetamido)-2-aminoethanesulfonic acid), Tricine (N-tris(hydroxymethyl)methylglycine), HEPES 4-2-hydroxyethyl-1-piperazineethanesulfonic acid), TES (2-{[tris(hydroxymethyl)methyl]amino}ethanesulfonic acid), MOPS (3-(N-morpholino)propanesulfonic acid), PIPES (piperazine-N,N'-bis(2-ethanesulfonic acid)), and MES (2-(N-morpholino)ethanesulfonic acid).

[0124] The pH of the reaction can influence the kinetics of the polymerase reaction, and can be used as one of the polymerase reaction conditions to obtain a reaction exhibiting two slow-step kinetics. The pH can be adjusted to a value that produces a two slow-step reaction mechanism. The pH is generally between about 6 and about 9. In some cases, the pH is between about 6.5 and about 8.0. For example, the pH can be by way of illustrations between about 6.5 and 7.5. The pH can be about 6.5, 6.6, 6.7, 6.8, 6.9, 7.0, 7.1, 7.2, 7.3, 7.4, or 7.5.

[0125] The temperature of the reaction can be adjusted in order to obtain a reaction exhibiting two slow-step kinetics. The reaction temperature may depend upon the type of polymerase which is employed. Temperatures between about 15° C and 90° C, between about 20° C and 50° C, between about 20° C and 40° C, or between about 20° C and 30° C can be used.

[0126] In one aspect, the present invention is directed to the use of a mixture of catalytic and non-catalytic metal ions in a nucleic acid synthesis reaction, to modulate the reaction kinetics of the complex. Thus, in at least one aspect, the invention is directed to nucleic acid synthesis reaction mixtures that include both catalytic and non-catalytic metals. The molar ratio of catalytic to non-catalytic metals in the reaction mixture will generally vary depending upon the type of kinetic modulation desired for a given synthesis reaction, where slower incorporation would suggest higher levels of non-catalytic metal ions. Typically, such ratios of catalytic to non-catalytic metals in the reaction mixture will vary from about 10:1 to

about 1:10, and preferably, from about 10:1 to about 1:5, depending upon the desired level of modulation, the particular enzyme system employed, the catalytic and non-catalytic metal cofactors that are used, and the reaction conditions. In particularly preferred aspects, the ratios of catalytic to non-catalytic metals will be in the range of from about 5:1 to about 1:1, with ratios of from about 2.5:1 to about 1.5:1 being particularly preferred.

[0127] In addition to the presence of such metals at the ratios described herein, the absolute concentration of such metals in the reaction mixtures will typically range from about 0.05 mM to about 50 mM, in some cases from about 0.1 mM to about 10 mM, in some cases from about 0.1 mM to about 5 mM. The composition can include, for example, from about 0.1 mM $MnCl_2$ to about 1 mM $MnCl_2$ and from about 0.1 mM $CaCl_2$ to about 2 mM $CaCl_2$; or from about 0.2 mM $MnCl_2$ to about 1 mM $MnCl_2$ and from about 0.4 mM $CaCl_2$ to about 1.5 mM $CaCl_2$.

[0128] As will be appreciated, the two slow-step enzyme systems described above can be used in combination with any other embodiments described herein, including: (1) reaction conditions that increase the frequency of branching fraction nonincorporation events; (2) modified recombinant polymerases that exhibit increased branching fractions as compared to the corresponding wild-type polymerases; (3) iterative sampling of unincorporatable nucleotides; (4) detection of noncognate branching events; (5) modified recombinant polymerases that exhibit altered translocation properties as compared to the corresponding wild-type polymerases; (6) modified recombinant polymerases that exhibit increased nucleotide residence time; (7) sequencing of nucleic acid templates using more than one mode; and/or (8) any other combination of embodiments described herein.

E. Enhanced Sequencing By Detection of Noncognate Branching Events

[0129] Nucleic acid sequencing approaches that utilize fluorescent-labeled nucleotide analogs typically require the detection of at least four colors – one color for each nucleotide representing the four different bases A, T/U, C, or G. Sequencing by incorporation methods, e.g., real time single molecule sequencing, also typically involves the detection of at least four different fluorescent labels corresponding to each of four nucleotide analogs. New approaches for reducing the number of different fluorescent labels to be detected by a sequencing system

are provided herein; these approaches can reduce the cost of nucleic acid sequencing by reducing, e.g., the number of color channels and other associated optical capabilities of the detection system.

[0130] The present invention provides new methods for nucleic acid sequencing that reduce the number of required colors to be detected by the sequencing system. Such methods utilize noncognate branching to determine which nucleotide is incorporated at a particular incorporation site during a nucleic acid polymerization reaction. For the purposes of this disclosure, noncognate branching refers to the branching (or iterative sampling) of a nucleotide, where the base of the nucleotide would incorrectly pair (not form a correct Watson-Crick base pair) with the base at the incorporation site of the template nucleic acid. For example, when the incorporation site of a template nucleic acid contains the base guanine (or G), it has been observed that branching (or sampling) of dGTP nucleotides or nucleotide analogs occurs at an appreciable frequency. The frequency of noncognate branching is sufficient to permit the identification of the nucleotide actually incorporated, regardless of whether the nucleotide actually incorporated is labeled. For instance, multiple signal pulses derived from noncognate branching (i.e., iterative sampling) of fluorescently-labeled G-containing nucleotides or nucleotide analogs can be utilized by a signal detection and sequencing system to determine that a C-containing nucleotide (which does not require a fluorescent label) was actually incorporated at that site. Thus, the temporal sequence of incorporation of four different nucleotides can be ascertained by utilizing, at most, three different fluorescent labels.

[0131] Noncognate branching of T-containing nucleotides has also been observed when the base at the incorporation site of the template is T. Accordingly, it is possible to determine the sequence of a nucleic acid template using a two-color detection system, where noncognate branching of G- and T-containing nucleotides – alone – permits the identification of incorporation of C- and A-containing nucleotides, thereby eliminating the requirement that C- and A-containing nucleotides be labeled. As will be appreciated, the present invention provides enhanced sequencing using noncognate branching, where the sampling noncognate nucleotide can be any nucleotide that would incorrectly pair with the corresponding base of the template nucleic acid.

[0132] It will be appreciated that noncognate branching events, in conjunction with a sequencing system that detects and accounts for noncognate branching events, are useful for sequencing applications other than those designed to reduce, e.g., the number of color channels required of the sequencing system. For example, even where sequencing is performed using four colors, noncognate branching can be used to validate or assist in the determination by a sequencing system of which nucleotide was incorporated at a particular site. Thus, when the base at the incorporation site of a template of unknown sequence is G, and the quantity or quality of signal pulses generated from sampling of dCTP analogs does not unambiguously permit assignment of dCTP as the incorporating nucleotide at that site, noncognate branching of labeled dGTP nucleotides facilitates the determination that dCTP was indeed incorporated at the site.

[0133] As will be appreciated, enhanced sequencing by detection and analysis of noncognate branching events can be used in combination with any other embodiments described herein, including: (1) reaction conditions that increase the frequency of branching fraction nonincorporation events; (2) modified recombinant polymerases that exhibit increased branching fractions as compared to the corresponding wild-type polymerases; (3) iterative sampling of unincorporatable nucleotides; (4) two slow-step enzyme systems; (5) modified recombinant polymerases that exhibit altered translocation properties as compared to the corresponding wild-type polymerases; (6) modified recombinant polymerases that exhibit increased nucleotide residence time; (7) sequencing of nucleic acid templates using more than one mode; and/or (8) any other combination of embodiments described herein.

II. SEQUENCING BY INCORPORATION USING MORE THAN ONE MODE OF SEQUENCING

[0134] A sequencing composition may be especially advantageous for one particular purpose or type of template nucleic acid, but may exhibit limitations for a second purpose or type of template nucleic acid. As described in detail herein, sequencing under conditions that promote a relatively high branching fraction, in combination with appropriate signal analysis, reduces the likelihood of base calling errors that result from undetected signal pulses generated from genuine nucleotide incorporation events. Accordingly, the base calling accuracy of

sequencing by incorporation reactions that exhibit a relatively high level of branch fraction nonincorporation events – when combined with a sequencing system that accounts for such redundant signal pulses per incorporation event – can be higher than the level of accuracy achieved under low branch rate conditions. On the other hand, sequencing under conditions in which the branching fraction is low typically results in improved sequence read length, as the number of incorporation events will be greater before the polymerase has an opportunity to dissociate from the template nucleic acid. Longer read lengths simplify assembly of contig information, e.g., to facilitate genomic sequencing.

[0135] De novo sequencing (sequencing a template nucleic acid of unknown sequence, e.g., genomic DNA of unknown sequence) is, in some cases, optimally performed under a combination of both high accuracy and high read length conditions. Identifying one particular sequencing composition that adequately fulfills both of these competing requirements is not necessary in the present invention. Instead, the present invention provides new methods and systems for sequencing a template nucleic acid, in which the template nucleic acid is sequenced using more than one sequencing mode ("multi-modal" or "variable mode" sequencing). As used herein, a sequencing mode refers to a sequencing composition (e.g., a mixture of a particular nucleic acid polymerase, nucleotides or nucleotide analogs, metal cofactors, and other components of a sequencing reaction) and other conditions that affect nucleic acid polymerization, e.g., reaction temperature. The methods and systems of the invention utilize a first sequencing mode that confers a particular benefit for purposes of generating a sequence readout (e.g., high accuracy), and then switch to a second (or more) sequencing mode that confers a benefit not realized during the first sequencing mode, e.g., a high sequence read length. As will be appreciated, any number of different modes can be employed by the methods and systems of the inventions, e.g., 2, 3, 4, 5, 10, or 20 different modes may be used to sequence a particular nucleic acid template until the desired results are achieved. For example, a template nucleic acid can be sequenced multiple times using a "high accuracy" mode and/or multiple times using a "high read length" mode, in order to obtain the accuracy and fold coverage useful for, e.g., de novo sequencing. Alternatively, a single template nucleic acid can be partially sequenced using, e.g., a high accuracy mode, and the mode can be switched prior to complete sequencing of the template to, e.g., a high read length mode.

[0136] A sequencing mode in which, e.g., high accuracy is desirable can employ, e.g., reaction conditions that induce a relatively high average branching fraction during the nucleic acid polymerization reaction. Such reaction conditions are described in detail herein. In one embodiment of the invention, one or more metal cofactors may be included in the sequencing composition at concentrations and/or ratios that induce a relatively high average branching fraction. Metal cofactors of the invention include, e.g., Mg^{++}, Mn^{++}, Zn^{++}, Co^{++}, Ca^{++}, Fe^{++}, Cr^{++}, and/or Sr^{++}. Reaction conditions in which metal cofactor identities, concentrations and/or ratios induce high levels of branching are described in detail herein. As described herein, a relatively high level of branching fraction nonincorporation events can be induced in a sequencing composition that includes, e.g., a relatively low concentration of Mn^{++}, a sequencing composition that includes Mg^{++}, a sequencing composition that includes both Mg^{++} and Mn^{++} with Mg^{++} being included at a higher concentration than Mn^{++}, and a sequencing composition that includes Mn^{++} and a metal cofactor other than Mg^{++}, e.g., calcium, barium, strontium, iron, cobalt, nickel, tin, zinc, and europium.

[0137] A sequencing mode in which, e.g., high read length is desirable can employ, e.g., reaction conditions that discourage branch fraction nonincorporation events during the nucleic acid polymerization reaction. Such reaction conditions are described herein. As described herein, reaction conditions that do not promote branching events include, e.g., Mn^{++} alone at a concentration at which branching is not induced (e.g., greater than 250 μM), or Mn^{++} in combination with a second metal cofactor, e.g., Mg^{++}, with the concentration of Mn^{++} being greater than the concentration of Mg^{++}. Desirably, sequencing modes of the present invention that produce high read lengths (e.g., a high number of successive incorporation events that are detected by sequencing systems of the invention) can produce read lengths of preferably more than 200 base pairs (bp) or more, more preferably 500 bp or more, more preferably 1000 bp or more, more preferably 10,000 bp or more, or more preferably 50,000 bp or more.

[0138] In order to permit the switching between a first sequencing mode and a second (or subsequent or more) sequencing mode, the sequencing composition – including the template nucleic acid and polymerase – is desirably confined within a structure to which additional components can be added to the sequencing composition to achieve the desired subsequent mode. In one aspect, for example, the sequencing of a template nucleic acid under

a first sequencing mode occurs within a structure to which a channel, e.g., a microfluidic channel, delivers the reagents necessary to achieve the subsequent mode. For example, switching from a high branch fraction sequencing mode, in which Mg^{++} is the sole catalytic metal ion, to a high read length sequencing mode can be achieved by delivery of Mn^{++} to the sequencing reaction, such that the final concentration of Mn^{++} is about 300 μM or more, about 400 μM or more, about 500 μM or more, about 700 μM or more, or about 1 mM or more.

[0139] In a preferred embodiment, the sequencing reactions take place in a structure that provides optical confinement, e.g., a nanohole or zero-mode waveguide. Further details regarding confinement strategies, substrates and systems for monitoring sequencing reactions can be found in co-pending published U.S. Patent Application No. 2007-0188750, and published International Patent Application No. WO 2007/095119, the full disclosures of which are incorporated herein by reference in their entirety for all purposes.

[0140] The sequencing methods and systems of the present invention can employ a wide variety of template nucleic acids. In some cases, the template nucleic acid is a linear template. For a linear template, switching between a high accuracy and a high read length mode can result in a read wherein a relatively long stretch of template nucleic acid is sequenced in a given time, wherein generally, multiple regions along the length will be sequenced with high accuracy. The knowledge that the high accuracy regions are arranged in a particular sequential manner can be useful in putting together the sequence of the nucleic acid. In some cases, the template nucleic acid is a circular template. For a circular template, switching between high accuracy and long read length modes can result in the same region of the template nucleic acid sequenced by each of the modes. In a preferred embodiment, the template to be sequenced in more than one sequencing mode is a single-stranded nucleic acid loop. Double-stranded templates can reanneal, reducing primer annealing efficiency and impeding the polymerase-catalyzed extension of a sequencing reaction. In fact, loops can also be preferable to linear templates because a DNA polymerase can only copy a linear template, e.g., to which a primer has been annealed, once before it falls off the distal end of the template. In contrast, a strand-displacing polymerase can replicate a contiguous nucleic acid loop several times. The primer that is annealed to the loop is eventually displaced at its 5'-end upon completion of one revolution of the polymerase around the nucleic acid loop, and as

polymerization and displacement continue, a linear, single-stranded product comprising several copies of the nucleic acid sequence of the loop is generated. Accordingly, using nucleic acid loops in sequencing can provide an internal sequencing control.

[0141] The methods for preparing closed, single-stranded nucleic acid loops include providing a genomic DNA, a cDNA, or a DNA concatamer and generating double-stranded fragments that each comprise a first strand (e.g., an exonuclease sensitive strand) and a second strand (e.g., an exonuclease resistant strand). In a following step, the two strands in each fragment are separated, and the resulting single-stranded fragments are circularized to produce closed single-stranded nucleic acid loops, which can then be used as templates in a high-throughput sequencing system. Further details regarding the preparation of single stranded nucleic acid loops, and their use as templates in high-throughput sequencing systems, can be found in Patel et al. PCT Application Serial No. PCT/US2009/001930 METHODS AND COMPOSITIONS FOR NUCLEIC ACID SAMPLE PREPARATION, incorporated herein by reference in its entirety for all purposes.

[0142] Systems for analyzing the data generated during sequencing of a template nucleic acid using more than one sequencing mode are also a feature of the invention. Such systems will include a signal detector, e.g., in the case of a plurality of arrayed sequencing reactions, an array detector, e.g., an EMCCD. The detector, is then operatively coupled to a data storage and processing system. In a first sequencing mode, e.g., a mode where a high branching fraction is induced, the processing system is capable of interpreting multiple, e.g.,. redundant, or iterative signal pulses for each nucleotide incorporation event during a sequencing reaction to call bases with increased accuracy. In a subsequent sequencing mode, e.g., a high read length mode in which little or no branching occurs, the processing system is capable of interpreting single signal pulses as nucleotide incorporation events. Further details regarding base calling during sequencing by incorporation methods are found in Tomaney et al. PCT Application Serial No. PCT/US2008/065996 METHODS AND PROCESSES FOR CALLING BASES IN SEQUENCING BY INCORPORATION METHODS, incorporated herein by reference in its entirety for all purposes.

[0143] The multi-modal sequencing method of the present invention optionally includes recording signals or signatures, e.g., in a computer readable medium. The signals or signatures can be stored, e.g., as graphic or digital information. Any typical recording device such as a hard drive, memory card, memory stick, optical storage device or floppy drive can be used to record detected signals. Signals or signal signatures can also be deconvoluted or translated to provide, e.g., sequence information, e.g., in sequencing systems of the invention. Signal processing equipment can include, e.g., a computer having appropriate software for converting signals into sequence or assay parameter information.

[0144] Signal detection optics can be coupled to cameras, digital processing apparatus, or the like, to record and analyze signals detected in the various systems herein, . Systems can include a microscope, a CCD, a phototube, a photodiode, an LCD, a scintillation counter, film for recording signals, and the like. A variety of commercially available peripheral equipment and software is available for digitizing, storing and analyzing a digitized video or digitized optical image, e.g., using PC (Intel x86 or pentium chip-compatible DOS™, OS2™ WINDOWS™, WINDOWS NT™ or WINDOWS95™ based machines), MACINTOSH™, LINUX, or UNIX based (e.g., SUN™ work station) computers or digital appliances. Computers and digital appliances can include software for analyzing and perfecting signal interpretation. This can include standard application software such as spreadsheet or database software for storing signal information. However, systems of the invention can also include statistical analysis software to interpret signal or signature data, e.g., to translate the data into nucleic acid sequence information. For example, many vendors, such as Partek Incorporated (St. Peters, Mo.; www.partek.com) provide software for pattern recognition which can be applied to signal interpretation and analysis. Algorithms for sequencing systems that can be adapted to the invention are also described in Tomaney et al. PCT Application Serial No. PCT/US2008/065996 METHODS AND PROCESSES FOR CALLING BASES IN SEQUENCING BY INCORPORATION METHODS, incorporated herein by reference in its entirety for all purposes. Once signal information has been converted into sequence information, standard sequence aalysis software can be used to assemble overlapping sequence information. For example, sequence contigs can be assembled using available software such as DNA Baser (Heracle Software, Germany), or Artemis 11 (Sanger Institute) "Artemis and ACT:

Viewing, annotating and comparing sequences stored in a relational database" Carver et al. Bioinformatics 2008 PMID: 18845581 DOI: 10.1093/bioinformatics/btn529).

[0145] Relationships between datasets (e.g., high accuracy versus high readlength data) can similarly be analyzed, e.g., by pattern recognition software, Bayes classifiers, neural networks, Monte Carlo analysis, Principal Component Analysis (PCA), etc. Further information regarding genetic algorithms and neural networks that can be used to analyze signal or signature information can be found in David E. Goldberg (1989) Genetic Algorithms in Search, Optimization and Machine Learning Addison-Wesley Pub Co; ISBN: 0201157675; Timothy Masters (1993) Practical Neural Network Recipes in C++(Book & Disk edition) Academic Pr; ISBN: 0124790402; Kevin Gurney (1999) An Introduction to Neural Networks, UCL Press, 1 Gunpowder Square, London EC4A 3DE, UK; Christopher M. Bishop (1995) Neural Networks for Pattern Recognition Oxford Univ Press; ISBN: 0198538642; Brian D. Ripley, N. L. Hjort (Contributor) (1995) Pattern Recognition and Neural Networks Cambridge Univ Pr (Short); ISBN: 0521460867;Rubinstein, R. Y.; Kroese, D. P. (2007) Simulation and the Monte Carlo Method (2nd ed.). New York: John Wiley & Sons. ISBN 9780470177938; Tarantola, Albert (2005) Inverse Problem Theory Philadelphia: Society for Industrial and Applied Mathematics ISBN 0898715725; Steeb (2008) The Nonlinear Workbook: Chaos, Fractals, Neural Networks, Genetic Algorithms, Gene Expression Programming, Support Vector Machine, Wavelets, Hidden Markov Models, Fuzzy Logic with C++, Java and SymbolicC++ Programs: 4th edition. World Scientific Publishing. ISBN 981-281-852-9; Sergios Theodoridis, Konstantinos Koutroumbas, (2009) Pattern Recognition (4th edition), Elsevier, ISBN 978-1-59749-272-0, and in a variety of other currently available references. Computers/digital appliances also optionally include or are operably coupled to user viewable display systems (monitors, CRTs, printouts, etc.), printers to print data relating to signal information, peripherals such as magnetic or optical storage drives, user input devices (keyboards, microphones, pointing devices) and the like.

[0146] One example of a system that utilizes more than one mode of sequencing is illustrated in **Figure 7**. As shown, system 700 includes substrate 702 of the invention, e.g., upon which sequencing reactions are performed. Optical energy source 704 and additional optical energy source 706 deliver excitation light to the substrate 702, via an optical train. As

shown, the optical train includes a dichroic mirror 708 that is transmissive to excitation wavelength of the light from laser 704, while being reflective of light of the wavelength produced by laser 706, allowing both illumination beams to be in the same path. The excitation light is then directed at a second dichroic mirror 710, that reflects the excitation light through objective lens 712, and onto the substrate 702. Optical signals from the substrate 702 are then collected by objective lens 712, and are passed through dichroic 710. The fluorescent signals are then subjected to spatial separation using, e.g., a dispersive optical element, such as an optical grating or prism 714. The separated signals are then focused upon array detector 716, e.g., an EMCCD. The detector, is then operatively coupled to a data storage and processing system, such as computer 718 for processing and storage of the signal data and presentation of the data in a user desired format, e.g., on printer 720.

[0147] As shown in Figure 7, in one example of multi-modal sequencing, a first sequencing mode (Panel I) is produced in which a sequencing by incorporation reaction occurs under higher branch fraction conditions than a second sequencing mode. During the first sequencing mode, the processing system of the computer is capable of interpreting multiple, e.g.,. redundant, or iterative signal pulses or signatures for each nucleotide incorporation event during a sequencing reaction to call bases. Subsequently, a second sequencing mode (Panel II) is produced in which a sequencing by incorporation reaction occurs with no or very few branch fraction nonincorporation events, and accordingly, longer readlengths. During the second sequencing mode, the processing system of the computer interprets one or very few signal pulses or signatures as a nucleotide incorporation event during the sequencing reaction to call bases. The system can utilize signal pulse or signature data from each sequencing mode in order to determine the sequence of the template nucleic acid. As will be appreciated, a portion of the nucleic acid template can be sequenced once (in one of the two or more sequencing modes) or two or more times (two or more times in the same or multiple sequencing modes).

[0148] Although illustrated as an optical train that is transmissive of fluorescent signals, e.g., as provided by dichroic **710**, it will be appreciated that fluorescence reflective systems may also be employed. Further details regarding base calling during sequencing by incorporation methods are found in Tomaney et al. PCT Application Serial No. PCT/US2008/065996 METHODS AND PROCESSES FOR CALLING BASES IN

SEQUENCING BY INCORPORATION METHODS, incorporated herein by reference in its entirety for all purposes.

[0149] As will be appreciated, sequencing of a nucleic acid template in more than one sequencing mode can be combined with (1) reaction conditions that increase the frequency of branching fraction nonincorporation events; (2) modified recombinant polymerases that exhibit increased branching fractions as compared to the corresponding wild-type polymerases; (3) iterative sampling of unincorporatable nucleotides; (4) two slow-step enzyme systems; (5) modified recombinant polymerases that exhibit altered translocation properties as compared to the corresponding wild-type polymerases; (6) modified recombinant polymerases that exhibit increased nucleotide residence time as compared to the corresponding wild-type polymerases; and/or (7) any other combination of embodiments described herein.

III. DELAYED TRANSLOCATION

[0150] The present invention also provides polymerases that exhibit delayed translocation as compared to parental/wild type enzymes. Time necessarily lapses between the incorporation of one nucleotide and the incorporation of the next nucleotide due to the sequential (rather than simultaneous) nature of nucleotide incorporation. The duration of this time lapse is determined primarily by the rate at which a polymerase translocates along a template polynucleotide between incorporation events ("translocation" refers to the movement of a DNA polymerase along a template polynucleotide from an initial enzyme binding site to a subsequent enzyme binding site, where the enzyme binding sites correlate to nucleotide incorporation sites). Upon incorporation of a nucleotide, a polymerase is unable to accept another nucleotide until it has gone through the translocation process and moved into the next incorporation site.

[0151] When a polymerase translocates at a typical wild-type rate from one incorporation site to the next, a single molecule sequencing (SMS) system has less time to identify and distinguish when the polymerase has moved into the next incorporation site because the rapid translocation affords little separation between the signal pulses arising from incorporation events at one incorporation site and the pulses arising from those events at the next incorporation site. Accordingly, SMS under conditions of typical polymerase

translocation rates can have an increased potential for deletion errors relative to the correct template sequence (i.e., signals from incorporation events may not be detected or may be detected but not interpreted as incorporation events). These deletions would be artifacts that constitute errors in sequencing reads.

[0152] In accordance with aspects of the present invention, however, the modified recombinant polymerases provided by the present invention that exhibit a translocation delay of longer duration as compared to parental (or wild-type) polymerases are advantageous in the context of SMS because they provide greater temporal separation (i.e., resolution) between the signal pulse emitted while the polymerase resides at one incorporation site from the pulse emitted once the polymerase has entered the next incorporation site. This increased translocation delay permits the sequencing system to interpret signal pulses as correlating to separate and distinct nucleotide incorporation events, and hence greatly diminishes the potential for missed pulses or inaccurately characterized pulses relative to the correct template sequence. Further, a delay in translocation enhances the ability of a sequencing system to distinguish branching events that exhibit short inter-pulse widths from incorporation events that, under the conditions of delayed translocation, exhibit markedly broader inter-pulse widths.

[0153] In one aspect, the polymerases of the present invention are engineered to exhibit two sequential translocation kinetic steps that occur at a slower rate as compared to a parental polymerase, e.g., a wild-type parental polymerase. The two translocation kinetic steps may occur at the same rate, or the ratio of their rates may vary up to about 1:5 or 5:1 or more. The occurrence of exceedingly short inter-pulse widths, e.g., the time delay between signal pulses derived from sequential nucleotide incorporation events, is markedly reduced when two slow translocation kinetic steps are engineered into the polymerase. However, the present invention also provides polymerases that exhibit a translocation delay arising from a slow translocation kinetic step in conjunction with a slow kinetic step not related to translocation, e.g., a slow nucleotide and/or nucleotide analog binding step. Further details regarding polymerase systems with two slow kinetic steps can be found in Bjornson et al. PCT Application Serial Number PCT/US2009/002003 TWO SLOW-STEP POLYMERASE ENZYME SYSTEMS AND METHODS, incorporated herein by reference in its entirety for all purposes.

[0154] The delay in translocation exhibited by the polymerases of the invention can be more than 2.5x greater, e.g., more than 5x greater, more than 10x greater, more than 15x greater, more than 50x greater, more than 100x greater, more than 1000x greater, or more than 10,000x greater as compared to a parental polymerase (e.g., a wild type Φ29 polymerase) or more. Modified recombinant polymerases of the invention can allow the translocation step of a template-dependent polymerization reaction to be observable, e.g., wherein an otherwise unobservable translocation step (e.g., due to a high rate of translocation of an unmodified polymerase and limitations of the detection system) becomes observable as a result of modification of the polymerase.

[0155] The present invention also provides modified recombinant polymerases that include a heterologous polypeptide sequence fused at or near the c-terminus of the polymerase and/or amino acid substitutions or deletions, in order to delay translocation of the polymerase. As noted above, when a polymerase translocates at a typical wild-type rate from one incorporation site to the next, a single molecule sequencing (SMS) system can fail to distinguish when the polymerase has moved into the next incorporation site because the rapid translocation affords little separation between the signal pulses arising from branching and incorporation events at one incorporation site and the pulses arising from those events at the next incorporation site. Accordingly, SMS under conditions of typical polymerase translocation rates is susceptible to incorrect insertions relative to the correct template sequence (i.e., multiple signal pulses arising from branching and incorporation events at a single incorporation site may be interpreted as pulses arising from incorporation events at more than one incorporation site). These insertions are artifacts that constitute errors in sequencing reads.

[0156] The present invention provides modified or recombinant DNA polymerases that exhibit decreased translocation rates in order to provide greater temporal separation (i.e., resolution) between the signal pulses emitted while the polymerase resides at one incorporation site from the pulses emitted once the polymerase has entered the next incorporation site. This decreased translocation rate permits the sequencing system to interpret clusters of signal pulses as correlating to nucleotide incorporation sites, and hence greatly diminishes the occurrence of incorrect insertions relative to the correct template sequence.

[0157] In one embodiment, delayed translocation is accomplished by encoding a Φ29 DNA polymerase with a heterologous polypeptide sequence fused to the c-terminus of the DNA polymerase. For the purposes of this disclosure, "heterologous" refers to a polypeptide sequence that is not present in the wild-type parental polymerase. That these polymerases retain their functionality is a surprising aspect of the invention. The active site of the polymerase is located in the c-terminal portion of the protein, and previous attempts to modify the c-terminal portion have rendered the polymerase inactive. In one aspect, the heterologous polypeptide sequence can include between 6 and 10 positively charged amino acids, e.g., histidine. This stretch of positively charged amino acids can be encoded immediately downstream of the c-terminus of the polymerase. An example polymerase of this embodiment, comprising a Φ29 polymerase fused to 10 histidine residues at its c-terminus (SEQ ID No. 1), exhibits a translocation rate of 84 sec^{-1}, which constitutes a 2.3x delay in translocation as compared to an unmodified parental Φ29 polymerase under identical conditions. In another aspect, an amino acid linker sequence, e.g., a Ser3Gly linker (e.g., Gly-Gly-Gly-Ser-Gly-Gly-Gly-Ser-Gly-Gly-Gly-Ser-Gly) is encoded between the stretch of positively charged amino acids and the C-terminus of the polymerase. An example polymerase of this embodiment, comprising a Φ29 polymerase fused to 10 histidine residues at its c-terminus, with a Ser3Gly linker between the polymerase and the histidine residues (SEQ ID No. 2), exhibits a translocation rate of 220 sec^{-1}, which constitutes a 2.4x delay in translocation as compared to an unmodified parental Φ29 polymerase under identical conditions. The affinity of positively charged amino acid residues at the c-terminus of the polymerase for the negatively charged phosphate backbone of DNA decreases the efficiency of polymerase translocation to the next nucleotide incorporation site, thereby causing a delay in translocation. Polymerases of the invention can also exhibit improved polymerase processivity, as the interaction between the polymerase and the phosphate backbone of DNA is enhanced by, e.g., the addition of positively-charged amino acid residues at the c-terminus.

[0158] In another embodiment, the polymerases that exhibit delayed translocation comprise amino acid substitutions and/or deletions that modulate interaction of the polymerases with the negatively-charged phosphate backbone of DNA. A number of specific examples of a modified polymerase, e.g. modified to delay polymerase translocation, are described herein. A

region of the polymerase responsible for interaction with the DNA template and primer is referred to herein as the binding cleft. Within the binding cleft are particular amino acid residues that interact with the phosphate backbone of DNA. The phosphate backbone constitutes a uniform interaction platform, e.g., nucleobase independent, and interactions with the phosphate backbone are altered in polymerases of the present invention, e.g., utilizing the negative charge of the backbone. For example, the interaction between a polymerase of the invention and the phosphate backbone of DNA can be enhanced by substituting neutral or negatively charged amino acids in the binding cleft with positively charged residues. However, mutations that delay polymerase translocation are not limited to this region of the polymerase. Relative to a wild-type Φ29 DNA polymerase, polymerase modifications of the present invention can include, e.g., any of the following mutations or combination of the following mutations: Asp570Lys; Asp570Ala; Asn313Lys; Asn313Ala; Gln303Lys; Gln303Ala; Gly532Ser; Met533delet; Cys530delet; Met533delet and Cys530delet; Gly532delet; Ala531Gly; Gly532Ser; Thr573Lys; Thr573Ala; Asn396Lys; Thr571Lys; Thr571Ala; Thr534Lys; Thr534Ala; Asp535Lys; Asp534Ala; Lys529Ala; and Lys529Asn. For the purposes of the present application, a mutation that includes a deletion at a particular residue position is presented by the amino acid abbreviation, followed by the residue position, followed by "del". Thus, for example, the Met533delet mutation presented above will be understood to mean a mutant polymerase in which the methionine at position 533 has been deleted.

[0159] The translocation delay exhibited by modified polymerases of the present invention, e.g., a modified Φ29 polymerase or a modified Φ 29-type polymerase, can be greater than the translocation delay exhibited by the parental polymerase or, e.g., about 2.5x greater, about 5x greater, about 10x greater, about 15x greater, about 50x greater or about 100x or more greater, as compared to a parental polymerase (e.g., a wild type Φ29 polymerase) under the standard conditions described above.

[0160] As will be appreciated, the above-identified modified or recombinant polymerases that display delayed translocation may optionally include additional modifications that confer other useful properties described herein, e.g., increased branching fraction, increased nucleotide analog residence time and/or increased processivity, etc. As will also be

appreciated, the above-identified modified or recombinant polymerases that display delayed translocation can be used in combination with any other embodiments described herein, including: (1) reaction conditions that increase the frequency of branching fraction nonincorporation events; (2) modified recombinant polymerases that exhibit increased branching fractions as compared to the corresponding wild-type polymerases; (3) iterative sampling of unincorporatable nucleotides; (4) two slow-step enzyme systems; (5) detection of noncognate branching events; (6) modified recombinant polymerases that exhibit increased nucleotide residence time; (7) sequencing of nucleic acid templates using more than one mode; and/or (8) any other combination of embodiments described herein.

IV. INCREASED RESIDENCE TIME

[0161] The present invention also provides modified recombinant polymerases that exhibit increased nucleotide or nucleotide analog residence time at an active site of the polymerase. During SMS, a number of signal parameters may be and generally are used for pulse identification. Two primary parameters are pulse intensity and pulse width, where pulse width relates to the duration of a signal pulse as detected by a signal detector of the sequencing system. Signal pulses that exhibit a short pulse width can arise from undesirable sources during SMS, e.g., transient pulses from labeled analogs in the reaction region that are not involved in an incorporation event. Incorporation events, on the other hand, are generally characterized by longer pulse durations stemming from increased residence time of the labeled nucleotide in the observation region by virtue of it being complexed with the polymerase.

[0162] Notwithstanding the differences in residence time duration upon incorporation, in some cases it would be desirable to further increase residence time for incorporated nucleotides, in order to further enhance the distinction between incorporation and transient signal events.

[0163] The modified recombinant polymerases of the present invention improve the accuracy of template-dependent polymerization reactions by increasing the residence times of nucleotides or nucleotide analogs at an active site of the polymerase, thereby producing signal pulses of increased width to further facilitate identification as signal pulses resulting from incorporation events.

[0164] As noted above, signal pulse intensity and width are primary parameters for signal pulse identification during template-dependent polymerization reactions, e.g., SMS utilizing dye-labeled nucleotide analogs. The signal pulse width is largely determined by the residence time of the nucleotide analog at the active site of the polymerase during the nucleotide incorporation step of the polymerase kinetic cycle. Insufficient residence times produce signal pulses of short duration and, accordingly, narrow pulse width that can prevent detection of genuine nucleotide incorporation events by a signal detector of the sequencing system. Undetected incorporation events arising from insufficient residence time and, accordingly, signal pulse width, constitute sequencing errors that diminish the reliability of SMS results.

[0165] The present invention provides modified or recombinant polymerases that improve sequence read accuracy by increasing the residence time of a nucleotide analog at the polymerase active site. During the polymerase kinetic cycle, interaction between a DNA polymerase and a nucleotide or nucleotide analog induces a conformational change in the polymerase, referred to herein as polymerase isomerization. During polymerase isomerization, the nucleotide or nucleotide analog is effectively immobilized at the active site of the polymerase. The duration of the isomerization step, therefore, affects the residence time of the nucleotide analog at the polymerase active site, and, accordingly, affects the signal pulse width as detected by a signal detector of a sequencing system. A slow isomerization step, therefore, can increase signal pulse width and enhance the accuracy of applications such as SMS. Further details regarding DNA polymerase enzymes and/or nucleotides or nucleotide analogs that provide altered residence times for enhanced nucleic acid sequence analysis and determination can be found in Rank et al. U.S. Application Serial No. 11/977,160 POLYMERASE ENZYMES AND REAGENTS FOR ENHANCED NUCELIC ACID SEQUENCING, incorporated herein by reference in its entirety for all purposes.

[0166] Completion of the nucleotide or nucleotide analog incorporation step of the DNA polymerase kinetic cycle requires dissociation of the polymerase from the incorporating nucleotide. Until dissociation occurs, the nucleotide, e.g., a dye-labeled nucleotide analog, is effectively immobilized at the active site of the polymerase. Slowing the dissociation step, therefore, results in increased nucleotide residence time at the active site of the polymerase and

increases the duration of a signal pulse emitted by a dye-labeled nucleotide analog as it is incorporated into the copy nucleic acid.

[0167] Polymerases of the present invention increase the residence time of nucleotides or nucleotide analogs by slowing the isomerization and/or dissociation steps of the polymerase kinetic cycle. Accordingly, these polymerases facilitate signal pulse detection by increasing signal pulse width. For example, a polymerase provided by the present invention comprises the amino acid substitution T368P and exhibits an average nucleotide residence time that is about 2x greater as compared to the residence time of a wild-type parental polymerase under identical conditions.

[0168] The nucleotide or nucleotide analog residence time exhibited by modified polymerases of the present invention, e.g., a modified Φ29 polymerase or a modified Φ 29-type polymerase, can be greater than the residence time exhibited by the parental polymerase or, e.g., about 1.5x greater, about 2.0x greater or about 3.0x or more greater, under the standard conditions described above.

[0169] As will be appreciated, the above-identified modified or recombinant polymerases that display increased nucleotide or nucleotide analog residence time may optionally include additional modifications that confer other useful properties described herein, e.g., increased branching fraction, delayed translocation and/or increased processivity. As will also be appreciated, the above-identified modified or recombinant polymerases that display increased nucleotide or nucleotide analog residence time can be used in combination with any other embodiments described herein, including: (1) reaction conditions that increase the frequency of branching fraction nonincorporation events; (2) modified recombinant polymerases that exhibit increased branching fractions as compared to the corresponding wild-type polymerases; (3) iterative sampling of unincorporatable nucleotides; (4) two slow-step enzyme systems; (5) detection of noncognate branching events; (6) modified recombinant polymerases that exhibit altered translocation properties as compared to the corresponding wild-type polymerases; (7) sequencing of nucleic acid templates using more than one mode; and/or (8) any other combination of embodiments described herein.

[0170] The properties of increased branching rates, delayed translocation and increased nucleotide or nucleotide analog residence time are particularly useful in the context of an incorporation of labeled nucleotides by the polymerase, e.g., as detected during sequencing by incorporation methods (including, e.g., SMS methods). For example, the invention provides, e.g., compositions that include one or more engineered or modified polymerase enzymes optionally with one or more template DNAs, and/or labeled or otherwise modified nucleotides or nucleotide analogs, where the composition exhibits increased branching rates and/or delayed polymerase translocation during template dependent polymerase-mediated nucleic acid synthesis. Methods, including SMS using these compositions, are also provided, as are general methods of making polymerases having the properties noted herein.

[0171] Accordingly, among other aspects, the present invention provides new polymerases that incorporate nucleotide analogs, such as phosphate labeled analogs, into a growing template copy during DNA amplification. These polymerases are modified such that they have increased branching rates and/or delayed translocation and/or increased residence time when incorporating the relevant analogs, and optionally have improved DNA-polymerase processivity as compared to corresponding wild-type parental polymerases (e.g., polymerases from which modified recombinant polymerases of the invention were derived, e.g., by mutation).

[0172] These new polymerases and reaction conditions are particularly well suited to DNA amplification and/or sequencing applications, particularly sequencing protocols that include detection in real time of the incorporation of labeled analogs into DNA amplicons, because the increased branching rate, delayed translocation and/or nucleotide residence time facilitates the correct determination of which labeled nucleotide is incorporated at a site during a template dependent polymerization reaction.

FURTHER DETAILS REGARDING SYSTEMS FOR SEQUENCING BY INCORPORATION

[0173] One example of a system for use in the present invention is illustrated in **Figure 1**. As shown, system 100 includes substrate 102 of the invention, e.g., upon which sequencing reactions are performed. Optical energy source 104 and additional optical energy source 106

deliver excitation light to the substrate 102, via an optical train. As shown, the optical train includes a dichroic mirror 108 that is transmissive to excitation wavelength of the light from laser 104, while being reflective of light of the wavelength produced by laser 106, allowing both illumination beams to be in the same path. The excitation light is then directed at a second dichroic mirror 110, that reflects the excitation light through objective lens 112, and onto the substrate 102. Optical signals from the substrate 102 are then collected by objective lens 112, and are passed through dichroic 110. The fluorescent signals are then subjected to spatial separation using, e.g., a dispersive optical element, such as an optical grating or prism 114. The separated signals are then focused upon array detector 116, e.g., an EMCCD. The detector, is then operatively coupled to a data storage and processing system, such as computer 118 for processing and storage of the signal data and presentation of the data in a user desired format, e.g., on printer 120. The processing system of the computer is capable of interpreting multiple, e.g.,. redundant, or iterative signal pulses for each nucleotide incorporation event during a sequencing reaction to call bases with increased accuracy. Although illustrated as an optical train that is transmissive of fluorescent signals, e.g., as provided by dichroic 110, it will be appreciated that fluorescence reflective systems may also be employed. Further details regarding base calling during sequencing by incorporation methods are found in Tomaney et al. PCT Application Serial No. PCT/US2008/065996 METHODS AND PROCESSES FOR CALLING BASES IN SEQUENCING BY INCORPORATION METHODS, incorporated herein by reference in its entirety for all purposes.

[0174] As will be appreciated, a number of other components may be included in the systems described herein, including optical filters for filtering background illumination or bleed-through illumination from the light sources, from the actual optical signals. Additionally, alternate optical trains may employ cascaded spectral filters in separating different spectral signal components. A monitor of the computer can display optical signal pulse patterns 122 generated under the conditions provided by the invention, e.g., increased branching, delayed polymerase translocation or increased nucleotide analog residence time. A variety of other optical configurations may additionally be employed in conjunction with the compositions of the invention.

[0175] In the context of nucleic acid sequencing methods, it will be appreciated that the signal sources each represent sequencing reactions, and particularly, polymerase-mediated, template-dependent primer extension reactions, where in preferred aspects, each base incorporation event results in a prolonged illumination (or localization) of one of four differentially labeled nucleotides being incorporated, so as to yield a recognizable pulse or signature that carries a distinguishable spectral profile or color.

[0176] In the context of the present invention, a signal event is first classified as to whether it constitutes a significant signal pulse or signature based upon whether such signal event meets any of a number of different criteria. Once identified or classified as a significant pulse or signature, the signal pulse or signature may be further assessed to determine whether the signal pulse or signature constitutes an incorporation event and may be called as a particular incorporated base. As will be appreciated, the basis for calling a particular signal event as a significant pulse or signature, and ultimately as an incorporation event, will be subject to a certain amount of error, based upon a variety of parameters as generally set forth herein. The reaction conditions and modified recombinant polymerases of the present invention diminish the error associated with base calling.

[0177] Once a particular signal is identified as a significant pulse or signature and is assigned a particular spectrum, e.g. color, the spectrally assigned pulse may be further assessed to determine whether the pulse or signature can be called an incorporation event and, as a result, call the base incorporated in the nascent strand, or its complement in the template sequence. Calling of bases from color assigned pulse or signature data will typically employ tests that identify the confidence level with which a base is called. Typically, such tests will take into account the data environment in which a signal was received, including a number of the same data parameters used in identifying significant pulses, etc. For example, such tests may include considerations of background signal levels, adjacent pulse signal parameters (spacing, intensity, duration, etc.), spectral image resolution, and a variety of other parameters. Such data may be used to assign a score to a given base call for a color assigned signal pulse or signature, where such scores are correlative of a probability that the base called is incorrect, e.g., 1 in 100 (99% accurate), 1 in 1000 (99.9% accurate), 1 in 10,000 (99.99% accurate), 1 in 100,000 (99.999% accurate), or even greater.

[0178] Once a base is called with sufficient accuracy, subsequent bases called in the same sequencing run, and in the same primer extension reaction, may then be appended to each previously called base to provide a sequence of bases in the overall sequence of the template or nascent strand. Iterative processing and further data processing, as described in greater detail below, can be used to fill in any blanks, correct any erroneously called bases, or the like for a given sequence.

POLYMERASES AND NUCLEOTIDE ANALOGS

[0179] Various polymerases may be used in the methods, compositions and systems described herein, including DNA polymerases, RNA polymerases, reverse transcriptases, and mutant or altered forms of any of the foregoing.

[0180] DNA polymerases that can be modified to increase the average branching fraction, decrease the translocation rate or increase nucleotide residence time are generally available. DNA polymerases are sometimes classified into six main groups based upon various phylogenetic relationships, e.g., with *E. coli* Pol I (class A), *E. coli* Pol II (class B), *E. coli* Pol III (class C), Euryarchaeotic Pol II (class D), human Pol beta (class X), and *E. coli* UmuC/DinB and eukaryotic RAD30/xeroderma pigmentosum variant (class Y). For a review of recent nomenclature, see, e.g., Burgers et al. (2001) "Eukaryotic DNA polymerases: proposal for a revised nomenclature" J Biol Chem. 276(47):43487-90. For a review of polymerases, see, e.g., Hübscher et al. (2002) "Eukaryotic DNA Polymerases" Annual Review of Biochemistry Vol. 71: 133-163; Alba (2001) "Protein Family Review: Replicative DNA Polymerases" Genome Biology 2(1):reviews 3002.1-3002.4; and Steitz (1999) "DNA polymerases: structural diversity and common mechanisms" J Biol Chem 274:17395-17398. The basic mechanisms of action for many polymerases have been determined. The sequences of literally hundreds of polymerases are publicly available, and the crystal structures for many of these have been determined, or can be inferred based upon similarity to solved crystal structures for homologous polymerases. For example, the crystal structure of Φ29, a preferred type of parental enzyme to be modified according to the invention, is available.

[0181] DNA polymerases and their properties are described in detail in, among other places, DNA Replication 2nd Edition, Komberg and Baker, W.H. Freeman, New York, N.Y.

(1991). Known conventional DNA polymerases useful in the invention include, but are not limited to, Pyrococcus furiosus (Pfu) DNA polymerase (Lundberg et al., 1991, Gene, 108: 1, Stratagene), Pyrococcus woesei (Pwo) DNA polymerase (Hinnisdaels et al., 1996, Biotechniques, 20:186-8, Boehringer Mannheim), Thermus thermophilus (Tth) DNA polymerase (Myers and Gelfand 1991, Biochemistry 30:7661), Bacillus stearothermophilus DNA polymerase (Stenesh and McGowan, 1977, Biochim Biophys Acta 475:32), Thermococcus litoralis (Tli) DNA polymerase (also referred to as Vent.TM. DNA polymerase, Cariello et al., 1991, Polynucleotides Res, 19: 4193, New England Biolabs), 9.degree. Nm.TM. DNA polymerase (New England Biolabs), Stoffel fragment, ThermoSequenase.RTM. (Amersham Pharmacia Biotech UK), Therminator.TM. (New England Biolabs), Thermotoga maritima (Tma) DNA polymerase (Diaz and Sabino, 1998 Braz J Med. Res, 31:1239), Thermus aquaticus (Taq) DNA polymerase (Chien et al., 1976, J. Bacteoriol, 127: 1550), DNA polymerase, Pyrococcus kodakaraensis KOD DNA polymerase (Takagi et al., 1997, Appl. Environ. Microbiol. 63:4504), JDF-3 DNA polymerase (from thermococcus sp. JDF-3, Patent application WO 0132887), Pyrococcus GB-D (PGB-D) DNA polymerase (also referred as Deep Vent.TM. DNA polymerase, Juncosa-Ginesta et al., 1994, Biotechniques, 16:820, New England Biolabs), UlTma DNA polymerase (from thermophile Thermotoga maritima; Diaz and Sabino, 1998 Braz J. Med. Res, 31:1239; PE Applied Biosystems), Tgo DNA polymerase (from thermococcus gorgonarius, Roche Molecular Biochemicals), E. coli DNA polymerase I (Lecomte and Doubleday, 1983, Polynucleotides Res. 11:7505), T7 DNA polymerase (Nordstrom et al., 1981, J Biol. Chem. 256:3112), and archaeal DP11/DP2 DNA polymerase II (Cann et al., 1998, Proc Natl Acad. Sci. USA 95:14250-5).

[0182] Reverse transcriptases useful in the invention include, but are not limited to, reverse transcriptases from HIV, HTLV-1, HTLV-II, FeLV, FIV, SIV, AMV, MMTV, MoMuLV and other retroviruses (see Levin, Cell 88:5-8 (1997); Verma, Biochim Biophys Acta. 473:1-38 (1977); Wu et al., CRC Crit Rev Biochem. 3:289-347 (1975)).

[0183] In preferred embodiments, the polymerases employed during the sequencing processes, and optionally during pre-sequencing synthesis, will typically possess strand-displacement activity to displace any primers downstream of the primer at which the strand synthesis is initiated. A preferred rolling circle polymerase exhibits strand-displacement

activity, and as such, a single circular template can be sequenced repeatedly to produce a sequence read comprising multiple copies of the complement of the template strand by displacing the nascent strand ahead of the translocating polymerase. Non-limiting examples of rolling circle polymerases suitable for the present invention include but are not limited to T5 DNA polymerase, T4 DNA polymerase holoenzyme, phage M2 DNA polymerase, phage PRD1 DNA polymerase, Klenow fragment of DNA polymerase, and certain polymerases that are modified or unmodified and chosen or derived from the phages Φ29, PRD1, Cp-1, Cp-5, Cp-7, Φ15, Φ1, Φ21, Φ25, BS 32 L17, PZE, PZA, Nf, M2Y (or M2), PR4, PR5, PR722, B103, SF5, GA-1, and related members of the *Podoviridae* family. In certain preferred embodiments, the polymerase is a modified Phi29 DNA polymerase, e.g., as described in U.S. Patent Publication No. 20080108082, incorporated herein by reference in its entirety for all purposes. Similarly, polymerases having enhanced activity for labeled nucleotides are also desirable. Examples of polymerase enzymes for use in various aspects of the invention include, e.g., those described in U.S. Patent Application Nos. 11/645,125, filed December 21, 2006; 11/645,135, filed December 21, 2006; 12/384,112, filed March 30, 2009; 61/094,843, filed September 5, 2008; and 61/072,645, filed March 31, 2008; as well as U.S. Patent Publication No. 20070196846 (the full disclosures of which are incorporated herein by reference in their entireties for all purposes), such as the E375Y/K512Y/T368F mutant polymerase described in the foregoing.

[0184] In addition to wild-type polymerases, chimeric polymerases made from a mosaic of different sources can be used. For example, Φ29 polymerases made taking sequences from more than one parental polymerase into account can be used as a starting point for mutating the polymerases of the invention. This can done using consideration of similarity regions between the polymerases to define consensus sequences that are used in the chimera, or can be done using gene shuffling technologies in which multiple Φ29-related polymerases are randomly or semi-randomly shuffled via available gene shuffling techniques (e.g., via "family gene shuffling"; see Crameri et al. (1998) "DNA shuffling of a family of genes from diverse species accelerates directed evolution" Nature 391:288–291; Clackson et al. (1991) "Making antibody fragments using phage display libraries" Nature 352:624-628; Gibbs et al. (2001) "Degenerate oligonucleotide gene shuffling (DOGS): a method for enhancing the frequency of

recombination with family shuffling" Gene 271:13-20; and Hiraga and Arnold (2003) "General method for sequence-independent site-directed chimeragenesis: J. Mol. Biol. 330:287-296). In these methods, the recombination points can be predetermined such that the gene fragments assemble in the correct order. However, the combinations, e.g., chimeras, can be formed at random. For example, using methods described in Clarkson, et al., 5 gene chimeras, e.g., comprising segments of a Phi29 polymerase, a PZA polymerase, a M2 polymerase, a B103 polymerase, and a GA-1 polymerase, with improved branching fractions can be generated.

[0185] Available DNA polymerase enzymes have also been modified in any of a variety of ways, e.g., to simplify production by making protease digested enzyme fragments such as the Klenow fragment recombinant, etc. As noted, polymerases have also been modified to confer improvements in specificity, processivity, and improved retention time of labeled nucleotides in polymerase-DNA-nucleotide complexes (e.g., WO 2007/076057 POLYMERASES FOR NUCLEOTIDE ANALOG INCORPORATION by Hanzel et al., and PCT/US2007/022459 POLYMERASE ENZYMES AND REAGENTS FOR ENHANCED NUCLEIC ACID SEQUENCING by Rank et al.) and to improve surface-immobilized enzyme activities (e.g., WO 2007/075987 ACTIVE SURFACE COUPLED POLYMERASES by Hanzel et al., and WO 2007/076057 PROTEIN ENGINEERING STRATEGIES TO OPTIMIZE ACTIVITY OF SURFACE ATTACHED PROTEINS by Hanzel et al.). Any of these available polymerases can be modified in accordance with the invention to increase the average branching fraction, delay translocation and/or increase nucleotide or nucleotide analog residence time.

[0186] The polymerase mutations and mutational strategies noted herein can be combined with available mutations and mutational strategies to confer additional improvements in, e.g., nucleotide analog specificity, enzyme processivity and the like. For example, the mutations and mutational strategies herein can be combined with those taught in, e.g., WO 2007/076057 POLYMERASES FOR NUCLEOTIDE ANALOG INCORPORATION by Hanzel et al., and PCT/US2007/022459 POLYMERASE ENZYMES AND REAGENTS FOR ENHANCED NUCLEIC ACID SEQUENCING by Rank et al. This combination of mutations/ mutational strategies can be used to impart several simultaneous improvements to a polymerase (decreased branch fraction formation, improved specificity, improved processivity, improved

retention time, improved stability of the closed complex, etc.). In addition, polymerases can be further modified for application-specific reasons, such as to improve activity of the enzyme when bound to a surface, as taught, e.g., in WO 2007/075987 ACTIVE SURFACE COUPLED POLYMERASES by Hanzel et al., and WO 2007/076057 PROTEIN ENGINEERING STRATEGIES TO OPTIMIZE ACTIVITY OF SURFACE ATTACHED PROTEINS by Hanzel et al., or to include purification or handling tags as is taught in the cited references and as is common in the art.

[0187] Specific mutations noted herein can be used alone or in combination with each other and/or with available mutations as described in the references noted above, or can be used in polymerases that lack such previously described mutations.

[0188] Many such polymerases that are suitable for modification are available, e.g., for use in sequencing, labeling and amplification technologies. For example, Human DNA Polymerase Beta is available from R&D systems. DNA polymerase I is available from Epicenter, GE Health Care, Invitrogen, New England Biolabs, Promega, Roche Applied Science, Sigma Aldrich and many others. The Klenow fragment of DNA Polymerase I is available in both recombinant and protease digested versions, from, e.g., Ambion, Chimerx, eEnzyme LLC, GE Health Care, Invitrogen, New England Biolabs, Promega, Roche Applied Science, Sigma Aldrich and many others. Φ29 DNA polymerase is available from e.g., Epicentre. Poly A polymerase, reverse transcriptase, Sequenase, SP6 DNA polymerase, T4 DNA polymerase, T7 DNA polymerase, and a variety of thermostable DNA polymerases (Taq, hot start, titanium Taq, etc.) are available from a variety of these and other sources. Recent commercial DNA polymerases include Phusion™ High-Fidelity DNA Polymerase, available from New England Biolabs; GoTaq® Flexi DNA Polymerase, available from Promega; RepliPHI™ Φ29 DNA Polymerase, available from Epicentre Biotechnologies; PfuUltra™ Hotstart DNA Polymerase, available from Stratagene; KOD HiFi DNA Polymerase, available from Novagen; and many others. Biocompare(dot)com provides comparisons of many different commercially available polymerases.

[0189] DNA polymerases that are preferred substrates for mutation to increase branch rates and/or decrease translocation rate include Taq polymerases, *E. coli* DNA Polymerase 1,

Klenow fragment, reverse transcriptases, Φ29 related polymerases including wild type Φ29 polymerase and derivatives of such polymerases such as exonuclease altered forms, T7 DNA polymerase, T5 DNA polymerase, an RB69 polymerase, etc.

[0190] In one aspect, the modified recombinant polymerases that exhibit increased branching fractions, delayed translocation and/or increased nucleotide or nucleotide analog residence time are Φ29-type DNA polymerases. For example, the modified recombinant DNA polymerases can be homologous to a wild-type Φ29 DNA polymerase, e.g., as described in U.S. Patent Nos. 5,001,050, 5,198,543, or 5,576,204. Alternately, the modified recombinant DNA polymerase can be homologous to other Φ29-type DNA polymerases, such as B103, GA-1, PZA, Φ15, BS32, M2Y, Nf, G1, Cp-1, PRD1, PZE, SF5, Cp-5, Cp-7, PR4, PR5, PR722, L17, Φ21, or the like. For nomenclature, see also, Meijer et al. (2001) "Φ29 Family of Phages" Microbiology and Molecular Biology Reviews, 65(2):261-287.

[0191] As discussed, various polymerases of the invention can incorporate one or more nucleotide analogs into a growing oligonucleotide chain. Upon incorporation, the analog can leave a residue that is the same as or different than a natural nucleotide in the growing oligonucleotide (the polymerase can incorporate any non-standard moiety of the analog, or can cleave it off during incorporation into the oligonucleotide). A "nucleotide analog" herein is a compound, that, in a particular application, functions in a manner similar or analogous to a naturally occurring nucleoside triphosphate (a "nucleotide"), and does not otherwise denote any particular structure. A nucleotide analog is an analog other than a standard naturally occurring nucleotide, i.e., other than A, G, C, T, or U, though upon incorporation into the oligonucleotide, the resulting residue in the oligonucleotide can be the same as (or different from) an A, G, C, T, or U residue.

[0192] Nucleotide analogs can also be modified to achieve any of the improved properties desired. For example, various linkers or other substituents can be incorporated into analogs that have the effect of altering the branching rate, residence time or improving processivity. Modifications to the analogs can include extending the phosphate chains, e.g., to include a hexa- or heptaphosphate group, and/or adding chemical linkers to extend the distance between the nucleotide base and the dye molecule, e.g., fluorescent dye molecule. As

described in detail herein, modifications to the analogs can include altering the analog such that the analog is unincorporatable. For example, nucleotide analogs of the invention can possess unhydrolyzable groups within the phosphate chain, such that the phosphoester linkage between the analog and the primer strand cannot be formed.

[0193] Many nucleotide analogs are available and can be incorporated by the polymerases of the invention. These include analog structures with core similarity to naturally occurring nucleotides, such as those that comprise one or more substituents on a phosphate, sugar or base moiety of the nucleoside or nucleotide relative to a naturally occurring nucleoside or nucleotide. In one embodiment, the nucleotide analog includes three phosphate containing groups; for example, the analog can be a labeled nucleoside triphosphate analog and/or an α-thiophosphate nucleotide analog having three phosphate groups. In one embodiment, a nucleotide analog can include one or more extra phosphate containing groups, relative to a nucleoside triphosphate. For example, a variety of nucleotide analogs that comprise, e.g., from 4-6 or more phosphates are described in detail in U.S. Patent Application No. 11/241,809, filed September 29, 2005, and incorporated herein by reference in its entirety for all purposes. Other exemplary useful analogs, including tetraphosphate and pentaphosphate analogs, are described in U.S. Patent 7,041,812, incorporated herein by reference in its entirety for all purposes.

[0194] Nucleotide analogs of the present invention may comprise any of a variety of detectable labels. Detectable labels generally denote a chemical moiety that provides a basis for detection of the analog compound separate and apart from the same compound lacking such a labeling group. Examples of labels include, e.g., optical labels, e.g., labels that impart a detectable optical property to the analog, electrochemical labels, e.g., labels that impart a detectable electrical or electrochemical property to the analog, and physical labels, e.g., labels that impart a different physical or spatial property to the analog, e.g., a mass tag or molecular volume tag. In some cases individual labels or combinations may be used that impart more than one of the aforementioned properties to the analogs of the invention.

[0195] Optionally, the labeling groups incorporated into the analogs comprise optically detectable moieties, such as luminescent, chemiluminescent, fluorescent, fluorogenic, chromophoric and/or chromogenic moieties, with fluorescent and/or fluorogenic labels being

preferred. A variety of different label moieties are readily employed in nucleotide analogs. Such groups include fluorescein labels, rhodamine labels, cyanine labels (i.e., Cy3, Cy5, and the like, generally available from the Amersham Biosciences division of GE Healthcare), the Alexa family of fluorescent dyes and other fluorescent and fluorogenic dyes available from Molecular Probes/Invitrogen, Inc. and described in 'The Handbook — A Guide to Fluorescent Probes and Labeling Technologies, Tenth Edition' (2005) (available from Invitrogen, Inc./Molecular Probes). A variety of other fluorescent and fluorogenic labels for use with nucleoside polyphosphates, and which would be applicable to the nucleotide analogs incorporated by the polymerases of the present invention, are described in, e.g., U.S. Patent Application Publication No. 2003/0124576, previously incorporated herein by reference in its entirety for all purposes.

[0196] Additional details regarding analogs and methods of making such analogs can be found in U.S. Patent Application No. 11/241,809, filed September 29, 2005, and incorporated herein by reference in its entirety for all purposes.

[0197] Thus, in one illustrative example, the analog can be a phosphate analog (e.g., an analog that has more than the typical number of phosphates found in nucleoside triphosphates) that include, e.g., an Alexa dye label. For example, an Alexa488 dye can be labeled on a delta phosphate of a tetraphosphate analog (denoted, e.g., A488dC4P or A488dA4P, for the Alexa488 labeled tetraphosphate analogs of C and A, respectively), or an Alexa568 or Alexa633 dye can be used (e.g., A568dC4P and A633dC4P, respectively, for labeled tetraphosphate analogs of C or A568dT6P for a labeled tetraphosphate analog of T), or an Alexa546 dye can be used (e.g., A546dG4P), or an Alexa594 dye can be used (e.g., A594dT4P). An Alexa555 dye (A555dC6P), an Alexa 647 dye (A647d6GP) and/or an Alexa660 dye (A660dA6P) can be used in, e.g., single molecule sequencing. Similarly, to facilitate color separation, a pair of fluorophores exhibiting FRET (fluorescence resonance energy transfer) can be labeled on a delta phosphate of a tetraphosphate analog (denoted, e.g., FAM-amb-A532dG4P or FAM-amb-A594dT4P).

[0198] Polymerases of the invention that exhibit branching phenotypes will display a branching fraction that is particular to the nucleotide analog included in the polymerization

reaction. For example, a polymerase of the invention may exhibit different branching fractions for A488dC4P than A488dC6P due, e.g., to the size differences between 4P and 6P analogs. It will be appreciated that polymerases of the present invention can be modified such that it exhibits the desired branching fraction for a particular nucleotide analog, e.g., a dye-labeled nucleotide analog with a particular number of phosphate groups.

APPLICATIONS FOR ENHANCED NUCLEIC ACID SEQUENCING

[0199] Polymerases of the invention, e.g., modified recombinant polymerases, are used in combination with nucleotides and/or nucleotide analogs, and nucleic acid templates (DNA or RNA) to copy the template nucleic acid. That is, a mixture of the polymerase, nucleotides/analogs, and optionally and other appropriate reagents, the template and a replication initiating moiety (e.g., primer) is reacted such that the polymerase synthesizes nucleic acid (e.g., extends the primer) in a template-dependent manner. The replication initiating moiety can be a standard oligonucleotide primer, or, alternatively, a component of the template, e.g., the template can be a self-priming single stranded DNA, a nicked double stranded DNA, or the like. Similarly, a terminal protein can serve as a initiating moiety. At least one nucleotide analog can be incorporated into the DNA. The template DNA can be a linear or circular DNA, and in certain applications, is desirably a circular template (e.g., for rolling circle replication or for sequencing of circular templates). Optionally, the composition can be present in an automated DNA replication and/or sequencing system.

[0200] Incorporation of labeled nucleotide analogs by the polymerases of the invention is particularly useful in a variety of different nucleic acid analyses, including real-time monitoring of DNA polymerization. The label can itself be incorporated, or more preferably, can be released during incorporation of the analog. For example, analog incorporation can be monitored in real-time by monitoring label release during incorporation of the analog by the polymerase. The portion of the analog that is incorporated can be the same as a natural nucleotide, or can include features of the analog that differ from a natural nucleotide.

[0201] In general, label incorporation or release can be used to indicate the presence and composition of a growing nucleic acid strand, e.g., providing evidence of template replication/amplification and/or sequence of the template. Signaling from the incorporation can

be the result of detecting labeling groups that are liberated from the incorporated analog, e.g., in a solid phase assay, or can arise upon the incorporation reaction. For example, in the case of FRET labels where a bound label is quenched and a free label is not, release of a label group from the incorporated analog can give rise to a fluorescent signal. Alternatively, the enzyme may be labeled with one member of a FRET pair proximal to the active site, and incorporation of an analog bearing the other member will allow energy transfer upon incorporation. The use of enzyme bound FRET components in nucleic acid sequencing applications is described, e.g., in U.S. Patent Application Publication No. 2003/0044781, incorporated herein by reference.

[0202] In one example reaction of interest, a polymerase reaction can be isolated within an extremely small observation volume that effectively results in observation of individual polymerase molecules. In a preferred aspect, such small observation volumes are provided by immobilizing the polymerase enzyme within a structural confinement, such as a nanohole (an aperture of less than 1 μM diameter through which a synthesis complex can be illuminated by optical energy) (See, e.g., co-pending Published U.S. Patent Application No. 2007-0188750, and published International Patent Application No. WO 2007/095119, the full disclosures of which are incorporated herein by reference in their entirety for all purposes) or nanoholes that additionally provide optical confinement, such as a zero-mode waveguide (ZMW). For a description of ZMWs and their application in single molecule analyses, and particularly nucleic acid sequencing, see, e.g., U.S. Patent Application Publication No. 2003/0044781, and U.S. Patent No. 6,917,726, each of which is incorporated herein by reference in its entirety for all purposes. See also Levene et al. (2003) "Zero-mode waveguides for single-molecule analysis at high concentrations" Science 299:682-686 and U.S. Patent Nos. 7,056,676, 7,056,661, 7,052,847, and 7,033,764, the full disclosures of which are incorporated herein by reference in their entirety for all purposes.

[0203] In general, a polymerase enzyme is complexed with the template strand in the presence of one or more nucleotides and/or one or more nucleotide analogs. For example, in certain embodiments, labeled analogs are present representing analogous compounds to each of the four natural nucleotides, A, T, G and C, e.g., in separate polymerase reactions, as in classical Sanger sequencing, or multiplexed together, e.g., in a single reaction, as in multiplexed sequencing approaches. When a particular base in the template strand is

encountered by the polymerase during the polymerization reaction, it complexes with an available analog that is complementary to such nucleotide, and incorporates that analog into the nascent and growing nucleic acid strand. In one aspect, incorporation can result in a label being released, e.g., in polyphosphate analogs, cleaving between the α and β phosphorus atoms in the analog, and consequently releasing the labeling group (or a portion thereof). The incorporation event is detected, either by virtue of a longer presence of the analog and, thus, the label, in the complex, or by virtue of release of the label group into the surrounding medium. Where different labeling groups are used for each of the types of analogs, e.g., A, T, G or C, identification of a label of an incorporated analog allows identification of that analog and consequently, determination of the complementary nucleotide in the template strand being processed at that time. Sequential reaction and monitoring permits a real-time monitoring of the polymerization reaction and determination of the sequence of the template nucleic acid. As noted above, in particularly preferred aspects, the polymerase enzyme/template complex is provided immobilized within an optical confinement that permits observation of an individual complex, e.g., a nanohole or zero mode waveguide.

[0204] In addition to their use in sequencing, the polymerases and/or reaction conditions of the invention are also useful in a variety of other genotyping analyses, e.g., SNP genotyping using single base extension methods, real time monitoring of amplification, e.g., RT-PCR methods, and the like. Further details regarding sequencing and nucleic acid amplification can be found, e.g., in Sambrook et al., Molecular Cloning - A Laboratory Manual (3rd Ed.), Vol. 1-3, Cold Spring Harbor Laboratory, Cold Spring Harbor, New York, 2000; Current Protocols in Molecular Biology, F.M. Ausubel et al., eds., Current Protocols, a joint venture between Greene Publishing Associates, Inc. and John Wiley & Sons, Inc., (supplemented through 2006); and PCR Protocols: A Guide to Methods and Applications (Innis *et al.* eds) Academic Press Inc. San Diego, CA (1990).

MOLECULAR MODELING-BASED MODIFICATION OF POLYMERASES TO INCREASE BRANCHING FRACTION, DELAY TRANSLOCATION OR INCREASE NUCLEOTIDE ANALOG RETENTION TIME

Structure-based design of recombinant polymerases

[0205] Structural data for a polymerase can be used to conveniently identify amino acid residues as candidates for mutagenesis to create recombinant polymerases, e.g., having modified active site regions that increase the branching fractions. For example, analysis of the three dimensional structure of a polymerase such as Φ29 can identify residues that are particularly relevant to branching, translocation and/or nucleotide residence time properties of the polymerase.

[0206] The three-dimensional structures of a large number of DNA polymerases have been determined by x-ray crystallography and nuclear magnetic resonance (NMR) spectroscopy, including the structures of polymerases with bound templates, nucleotides, and/or nucleotide analogs. Many such structures are freely available for download from the Protein Data Bank, at (www(dot)rcsb(dot)org/pdb. Structures, along with domain and homology information, are also freely available for search and download from the National Center for Biotechnology Information's Molecular Modeling DataBase, at www(dot)ncbi(dot)nlm(dot)nih(dot)gov/Structure/MMDB/mmdb(dot)shtml. The structures of additional polymerases can be modeled, for example, based on homology of the polymerases with polymerases whose structures have already been determined. Alternatively, the structure of a given polymerase, optionally complexed with a template and/or nucleotide analog, or the like, can be determined.

[0207] Techniques for crystal structure determination are well known. See, for example, McPherson (1999) Crystallization of Biological Macromolecules Cold Spring Harbor Laboratory; Bergfors (1999) Protein Crystallization International University Line; Mullin (1993) Crystallization Butterwoth-Heinemann; Stout and Jensen (1989) X-ray structure determination: a practical guide, 2nd Edition Wiley Publishers, New York; Ladd and Palmer (1993) Structure determination by X-ray crystallography, 3rd Edition Plenum Press, NewYork; Blundell and Johnson (1976) Protein Crystallography Academic Press, New York; Glusker and Trueblood (1985) Crystal structure analysis: A primer, 2nd Ed. Oxford University Press, NewYork; International Tables for Crystallography, Vol. F. Crystallography of Biological Macromolecules; McPherson (2002) Introduction to Macromolecular Crystallography Wiley-Liss; McRee and David (1999) Practical Protein Crystallography, Second Edition Academic Press; Drenth (1999) Principles of Protein X-Ray Crystallography (Springer Advanced Texts in

Chemistry) Springer-Verlag; Fanchon and Hendrickson (1991) Chapter 15 of Crystallographic Computing, Volume 5 IUCr/Oxford University Press; Murthy (1996) Chapter 5 of Crystallographic Methods and Protocols Humana Press; Dauter et al. (2000) "Novel approach to phasing proteins: derivatization by short cryo-soaking with halides" Acta Cryst.D56:232-237; Dauter (2002) "New approaches to high-throughput phasing" Curr. Opin. Structural Biol. 12:674–678; Chen et al. (1991) "Crystal structure of a bovine neurophysin-II dipeptide complex at 2.8 Å determined from the single-wavelength anomalous scattering signal of an incorporated iodine atom" Proc. Natl Acad. Sci. USA, 88:4240–4244; and Gavira et al. (2002) "*Ab initio* crystallographic structure determination of insulin from protein to electron density without crystal handling" Acta Cryst.D58:1147-1154.

[0208] In addition, a variety of programs to facilitate data collection, phase determination, model building and refinement, and the like are publicly available. Examples include, but are not limited to, the HKL2000 package (Otwinowski and Minor (1997) "Processing of X-ray Diffraction Data Collected in Oscillation Mode" Methods in Enzymology 276:307-326), the CCP4 package (Collaborative Computational Project (1994) "The CCP4 suite: programs for protein crystallography" Acta Crystallogr D 50:760-763), SOLVE and RESOLVE (Terwilliger and Berendzen (1999) Acta Crystallogr D 55 (Pt 4):849-861), SHELXS and SHELXD (Schneider and Sheldrick (2002) "Substructure solution with SHELXD" Acta Crystallogr D Biol Crystallogr 58:1772-1779), Refmac5 (Murshudov et al. (1997) "Refinement of Macromolecular Structures by the Maximum-Likelihood Method" Acta Crystallogr D 53:240-255), PRODRG (van Aalten et al. (1996) "PRODRG, a program for generating molecular topologies and unique molecular descriptors from coordinates of small molecules" J Comput Aided Mol Des 10:255-262), and O (Jones et al. (1991) "Improved methods for building protein models in electron density maps and the location of errors in these models" Acta Crystallogr A 47 (Pt 2):110-119).

[0209] Techniques for structure determination by NMR spectroscopy are similarly well described in the literature. See, e.g., Cavanagh et al. (1995) Protein NMR Spectroscopy: Principles and Practice, Academic Press; Levitt (2001) Spin Dynamics: Basics of Nuclear Magnetic Resonance, John Wiley & Sons; Evans (1995) Biomolecular NMR Spectroscopy, Oxford University Press; Wüthrich (1986) NMR of Proteins and Nucleic Acids (Baker Lecture

Series), Kurt Wiley-Interscience; Neuhaus and Williamson (2000) The Nuclear Overhauser Effect in Structural and Conformational Analysis, 2nd Edition, Wiley-VCH; Macomber (1998) A Complete Introduction to Modern NMR Spectroscopy, Wiley-Interscience; Downing (2004) Protein NMR Techniques (Methods in Molecular Biology), 2nd edition, Humana Press; Clore and Gronenborn (1994) NMR of Proteins (Topics in Molecular and Structural Biology), CRC Press; Reid (1997) Protein NMR Techniques, Humana Press; Krishna and Berliner (2003) Protein NMR for the Millenium (Biological Magnetic Resonance), Kluwer Academic Publishers; Kiihne and De Groot (2001) Perspectives on Solid State NMR in Biology (Focus on Structural Biology, 1), Kluwer Academic Publishers; Jones et al. (1993) Spectroscopic Methods and Analyses: NMR, Mass Spectrometry, and Related Techniques (Methods in Molecular Biology, Vol. 17), Humana Press; Goto and Kay (2000) Curr. Opin. Struct. Biol. 10:585; Gardner (1998) Annu. Rev. Biophys. Biomol. Struct. 27:357; Wüthrich (2003) Angew. Chem. Int. Ed. 42:3340; Bax (1994) Curr. Opin. Struct. Biol. 4:738; Pervushin et al. (1997) Proc. Natl. Acad. Sci. U.S.A. 94:12366; Fiaux et al. (2002) Nature 418:207; Fernandez and Wider (2003) Curr. Opin. Struct. Biol. 13:570; Ellman et al. (1992) J. Am. Chem. Soc. 114:7959; Wider (2000) BioTechniques 29:1278-1294; Pellecchia et al. (2002) Nature Rev. Drug Discov. (2002) 1:211-219; Arora and Tamm (2001) Curr. Opin. Struct. Biol. 11:540-547; Flaux et al. (2002) Nature 418:207-211; Pellecchia et al. (2001) J. Am. Chem. Soc. 123:4633-4634; and Pervushin et al. (1997) Proc. Natl. Acad. Sci. USA 94:12366-12371.

[0210] The structure of a polymerase, or polymerase bound to a DNA or with a given nucleotide analog incorporated into the active site can, as noted, be directly determined, e.g., by x-ray crystallography or NMR spectroscopy, or the structure can be modeled based on the structure of the polymerase and/or a structure of a polymerase with a natural nucleotide bound. The active site or other relevant domain of the polymerase can be identified, for example, by homology with other polymerases, examination of polymerase-template or polymerase-nucleotide co-complexes, biochemical analysis of mutant polymerases, and/or the like. The position of a nucleotide analog (as opposed to an available nucleotide structure) in the active site can be modeled, for example, by projecting the location of non-natural features of the analog (e.g., additional phosphate or phosphonate groups in the phosphorus containing chain linked to the nucleotide, e.g., tetra, penta or hexa phosphate groups, detectable labeling groups,

e.g., fluorescent dyes, or the like) based on the previously determined location of another nucleotide or nucleotide analog in the active site.

[0211] Such modeling of the nucleotide analog or template (or both) in the active site can involve simple visual inspection of a model of the polymerase, for example, using molecular graphics software such as the PyMOL viewer (open source, freely available on the World Wide Web at www(dot)pymol(dot)org) or Insight II (commercially available from Accelrys at (www (dot) accelrys (dot) com/products/insight). Alternatively, modeling of the active site complex of the polymerase or a putative mutant polymerase, for example, can involve computer-assisted docking, molecular dynamics, free energy minimization, and/or like calculations. Such modeling techniques have been well described in the literature; see, e.g., Babine and Abdel-Meguid (eds.) (2004) Protein Crystallography in Drug Design, Wiley-VCH, Weinheim; Lyne (2002) "Structure-based virtual screening: An overview" Drug Discov. Today 7:1047-1055; Molecular Modeling for Beginners, at (www (dot) usm (dot) maine (dot) edu/~rhodes/SPVTut/index (dot) html; and Methods for Protein Simulations and Drug Design at (www (dot) dddc (dot) ac (dot) cn/embo04; and references therein. Software to facilitate such modeling is widely available, for example, the CHARMm simulation package, available academically from Harvard University or commercially from Accelrys (at www (dot) accelrys (dot) com), the Discover simulation package (included in Insight II, *supra*), and Dynama (available at (www(dot) cs (dot) gsu (dot) edu/~cscrwh/progs/progs (dot) html). See also an extensive list of modeling software at (www (dot) netsci (dot) org/Resources/Software/Modeling/MMMD/top (dot) html.

[0212] Visual inspection and/or computational analysis of a polymerase model can identify relevant features of the active site or other domain, including, for example, amino acid residues of domains that are in close proximity to one another (to stabilize inter-domain interactions) residues in the active site that interact with the nucleotide or analog, or that modulate how large a binding pocket for the analog is relative to the analog. That is, inter-domain amino acid contacts can stabilize the closed complex, and/or the size or composition (e.g., position of charged or hydrophobic residues) of the binding pocket in the active site can control entry and release of the nucleotide, which can affect branching rate. A residue can, for example, be deleted or replaced with a residue having a different (smaller, larger, ionic, non-

ionic, etc.) side chain. Similarly, residues that can be altered to introduce desirable interactions with the nucleotide analog can be identified to reduce branching. Such a residue can be replaced with a residue that is complementary with, e.g., a non-natural feature of the analog, for example, with a residue that can hydrogen bond to the analog (e.g., serine, threonine, histidine, asparagine, or glutamine), a hydrophobic residue that can interact with a hydrophobic group on the analog, an aromatic residue that can provide favorable hydrophobic interactions with a group on the analog (e.g., a fluorophore), an aromatic residue that can engage in a π-π or edge-face stacking interaction with an aromatic group in the analog, a residue that can engage in a cation-π interaction with the analog, or a charged residue (e.g., aspartic or glutamic acid, or lysine, arginine, or histidine) that can electrostatically interact with an oppositely charged moiety on the analog (e.g., an additional phosphate group).

[0213] Thus, in addition to methods of using the polymerases and other compositions herein, the present invention also includes methods of making the polymerases. As described, methods of making a recombinant DNA polymerase can include structurally modeling a first polymerase, e.g., using any available crystal structure and molecular modeling software or system. Based on the modeling, one or more feature affecting closed complex stability, or nucleotide access or removal to or from the active site (and, thereby, branching) and/or binding of a DNA or nucleotide analog within the active site region is identified. These residues can be, e.g., in the active site, an exonuclease, TPR2 or thumb domain (or interface between domains) or proximal to such domains. The DNA polymerase is mutated to include non-natural residues at such positions, and then screened for an activity of interest.

Mutating Polymerases

[0214] Various types of mutagenesis are optionally used in the present invention, e.g., to modify polymerases to produce variants, e.g., in accordance with polymerase models and model predictions as discussed above, or using random or semi-random mutational approaches. In general, any available mutagenesis procedure can be used for making polymerase mutants. Such mutagenesis procedures optionally include selection of mutant nucleic acids and polypeptides for one or more activity of interest (e.g., decreased branch fraction, increased or decreased complex stability, improved processivity, and/or improved k_{off}, K_m, V_{max}, k_{cat} etc.,

e.g., for a given nucleotide analog). Procedures that can be used include, but are not limited to: site-directed point mutagenesis, random point mutagenesis, in vitro or in vivo homologous recombination (DNA shuffling and combinatorial overlap PCR), mutagenesis using uracil containing templates, oligonucleotide-directed mutagenesis, phosphorothioate-modified DNA mutagenesis, mutagenesis using gapped duplex DNA, point mismatch repair, mutagenesis using repair-deficient host strains, restriction-selection and restriction-purification, deletion mutagenesis, mutagenesis by total gene synthesis, degenerate PCR, double-strand break repair, and many others known to persons of skill. The starting polymerase for mutation can be any of those noted herein, including available polymerase mutants such as those identified e.g., in WO 2007/076057 POLYMERASES FOR NUCLEOTIDE ANALOG INCORPORATION by Hanzel et al.; PCT/US2007/022459 POLYMERASE ENZYMES AND REAGENTS FOR ENHANCED NUCLEIC ACID SEQUENCING; Hanzel et al. WO 2007/075987 ACTIVE SURFACE COUPLED POLYMERASES; and Hanzel et al. WO 2007/076057 PROTEIN ENGINEERING STRATEGIES TO OPTIMIZE ACTIVITY OF SURFACE ATTACHED PROTEINS.

[0215] Optionally, mutagenesis can be guided by known information from a naturally occurring polymerase molecule, or of a known altered or mutated polymerase (e.g., using an existing mutant polymerase as noted in the preceding references), e.g., sequence, sequence comparisons, physical properties, crystal structure and/or the like as discussed above. However, in another class of embodiments, modification can be essentially random (e.g., as in classical or "family" DNA shuffling, see, e.g., Crameri et al. (1998) "DNA shuffling of a family of genes from diverse species accelerates directed evolution" Nature 391:288–291).

[0216] Additional information on mutation formats is found in: Sambrook et al., Molecular Cloning - A Laboratory Manual (3rd Ed.), Vol. 1-3, Cold Spring Harbor Laboratory, Cold Spring Harbor, New York, 2000 ("Sambrook"); Current Protocols in Molecular Biology, F.M. Ausubel et al., eds., Current Protocols, a joint venture between Greene Publishing Associates, Inc. and John Wiley & Sons, Inc., (supplemented through 2006) ("Ausubel")) and PCR Protocols A Guide to Methods and Applications (Innis *et al.* eds) Academic Press Inc. San Diego, CA (1990) (Innis). The following publications and references cited within provide additional detail on mutation formats: Arnold, *Protein engineering for unusual environments*,

Current Opinion in Biotechnology 4:450-455 (1993); Bass et al., *Mutant Trp repressors with new DNA-binding specificities*, Science 242:240-245 (1988); Bordo and Argos (1991) *Suggestions for "safe Residue Substitutions in Site-directed Mutagenesis* 217:721-729; Botstein & Shortle, *Strategies and applications of in vitro mutagenesis*, Science 229:1193-1201(1985); Carter et al., *Improved oligonucleotide site-directed mutagenesis using M13 vectors*, Nucl. Acids Res. 13: 4431-4443 (1985); Carter, *Site-directed mutagenesis*, Biochem. J. 237:1-7 (1986); Carter, *Improved oligonucleotide-directed mutagenesis using M13 vectors*, Methods in Enzymol. 154: 382-403 (1987); Dale et al., *Oligonucleotide-directed random mutagenesis using the phosphorothioate method*, Methods Mol. Biol. 57:369-374 (1996); Eghtedarzadeh & Henikoff, *Use of oligonucleotides to generate large deletions*, Nucl. Acids Res. 14: 5115 (1986); Fritz et al., *Oligonucleotide-directed construction of mutations: a gapped duplex DNA procedure without enzymatic reactions in vitro*, Nucl. Acids Res. 16: 6987-6999 (1988); Grundström et al., *Oligonucleotide-directed mutagenesis by microscale 'shot-gun' gene synthesis*, Nucl. Acids Res. 13: 3305-3316 (1985); Hayes (2002) *Combining Computational and Experimental Screening for rapid Optimization of Protein Properties* PNAS 99(25) 15926-15931; Kunkel, *The efficiency of oligonucleotide directed mutagenesis*, in Nucleic Acids & Molecular Biology (Eckstein, F. and Lilley, D.M.J. eds., Springer Verlag, Berlin)) (1987); Kunkel, *Rapid and efficient site-specific mutagenesis without phenotypic selection*, Proc. Natl. Acad. Sci. USA 82:488-492 (1985); Kunkel et al., *Rapid and efficient site-specific mutagenesis without phenotypic selection*, Methods in Enzymol. 154, 367-382 (1987); Kramer et al., *The gapped duplex DNA approach to oligonucleotide-directed mutation construction*, Nucl. Acids Res. 12: 9441-9456 (1984); Kramer & Fritz *Oligonucleotide-directed construction of mutations via gapped duplex DNA*, Methods in Enzymol. 154:350-367 (1987); Kramer et al., *Point Mismatch Repair*, Cell 38:879-887 (1984); Kramer et al., *Improved enzymatic in vitro reactions in the gapped duplex DNA approach to oligonucleotide-directed construction of mutations*, Nucl. Acids Res. 16: 7207 (1988); Ling et al., *Approaches to DNA mutagenesis: an overview*, Anal Biochem. 254(2): 157-178 (1997); Lorimer and Pastan Nucleic Acids Res. 23, 3067-8 (1995); Mandecki, *Oligonucleotide-directed double-strand break repair in plasmids of Escherichia coli: a method for site-specific mutagenesis*, Proc. Natl. Acad. Sci. USA, 83:7177-7181 (1986); Nakamaye & Eckstein, *Inhibition of restriction endonuclease Nci I cleavage by*

phosphorothioate groups and its application to oligonucleotide-directed mutagenesis, Nucl. Acids Res. 14: 9679-9698 (1986); Nambiar et al., *Total synthesis and cloning of a gene coding for the ribonuclease S protein*, Science 223: 1299-1301 (1984); Sakamar and Khorana, *Total synthesis and expression of a gene for the a-subunit of bovine rod outer segment guanine nucleotide-binding protein (transducin)*, Nucl. Acids Res. 14: 6361-6372 (1988); Sayers et al., *Y-T Exonucleases in phosphorothioate-based oligonucleotide-directed mutagenesis*, Nucl. Acids Res. 16:791-802 (1988); Sayers et al., *Strand specific cleavage of phosphorothioate-containing DNA by reaction with restriction endonucleases in the presence of ethidium bromide*, (1988) Nucl. Acids Res. 16: 803-814; Sieber, et al., Nature Biotechnology, 19:456-460 (2001); Smith, *In vitro mutagenesis*, Ann. Rev. Genet. 19:423-462(1985); Methods in Enzymol. 100: 468-500 (1983); Methods in Enzymol. 154: 329-350 (1987); Stemmer, Nature 370, 389-91 (1994); Taylor et al., *The use of phosphorothioate-modified DNA in restriction enzyme reactions to prepare nicked DNA*, Nucl. Acids Res. 13: 8749-8764 (1985); Taylor et al., *The rapid generation of oligonucleotide-directed mutations at high frequency using phosphorothioate-modified DNA*, Nucl. Acids Res. 13: 8765-8787 (1985); Wells et al., *Importance of hydrogen-bond formation in stabilizing the transition state of subtilisin*, Phil. Trans. R. Soc. Lond. A 317: 415-423 (1986); Wells et al., *Cassette mutagenesis: an efficient method for generation of multiple mutations at defined sites*, Gene 34:315-323 (1985); Zoller & Smith, *Oligonucleotide-directed mutagenesis using M13-derived vectors: an efficient and general procedure for the production of point mutations in any DNA fragment*, Nucleic Acids Res. 10:6487-6500 (1982); Zoller & Smith, *Oligonucleotide-directed mutagenesis of DNA fragments cloned into M13 vectors*, Methods in Enzymol. 100:468-500 (1983); Zoller & Smith, *Oligonucleotide-directed mutagenesis: a simple method using two oligonucleotide primers and a single-stranded DNA template*, Methods in Enzymol. 154:329-350 (1987); Clackson et al. (1991) "Making antibody fragments using phage display libraries" Nature 352:624-628; Gibbs et al. (2001) "Degenerate oligonucleotide gene shuffling (DOGS): a method for enhancing the frequency of recombination with family shuffling" Gene 271:13-20; and Hiraga and Arnold (2003) "General method for sequence-independent site-directed chimeragenesis: J. Mol. Biol. 330:287-296. Additional details on many of the above methods can be found in Methods in

Enzymology Volume 154, which also describes useful controls for trouble-shooting problems with various mutagenesis methods.

Determining Kinetic Parameters

[0217] The polymerases of the invention can be screened or otherwise tested to determine whether the polymerase displays a modified activity for or with a nucleotide analog or template as compared to the first DNA polymerase (e.g., a corresponding wild-type or available mutant polymerase from which the recombinant polymerase of the invention was derived). For example, branch fraction, k_{off}, k_{cat}, K_m, V_{max}, k_{cat}/K_m, V_{max}/K_m, k_{pol}, and/or K_d of the recombinant DNA polymerase for the nucleotide (or analog) or template nucleic acid can be determined. The enzyme perfection metric k_{cat}/K_m is also a useful measure, e.g., for assessing branch rate. k_{cat}/K_m is a measure of substrate binding that leads to product formation (and, thus, includes terms defining binding K_d and inversely predicts branching fraction formation).

[0218] As is well-known in the art, for enzymes obeying simple Michaelis-Menten kinetics, kinetic parameters are readily derived from rates of catalysis measured at different substrate concentrations. The Michaelis-Menten equation, $V=V_{max}[S]([S]+K_m)^{-1}$, relates the concentration of uncombined substrate ([S], approximated by the total substrate concentration), the maximal rate (V_{max}, attained when the enzyme is saturated with substrate), and the Michaelis constant (K_m, equal to the substrate concentration at which the reaction rate is half of its maximal value), to the reaction rate (V). Details regarding K_{off} determination are described above. In general, the dissociation rate can be measured in any manner that detects the polymerase/ DNA complex over time. This includes stopped flow spectroscopy, or even simply by taking aliquots over time and testing for polymerase activity on the template of interest. Free polymerase is captured with a polymerase trap after dissociation, e.g., by incubation in the presence of heparin or an excess of competitor DNA (e.g., non-specific salmon sperm DNA, or the like).

[0219] For many enzymes, K_m is equal to the dissociation constant of the enzyme-substrate complex and is thus a measure of the strength of the enzyme-substrate complex. For such an enzyme, in a comparison of K_ms, a lower K_m represents a complex with stronger binding, while a higher Km represents a complex with weaker binding. The ratio k_{cat}/K_m,

sometimes called the specificity constant, represents the apparent rate constant for combination of substrate with free enzyme. The larger the specificity constant, the more efficient the enzyme is in binding the substrate and converting it to product (this provides an inverse measure of branching rate, as branching rate is the rate at which the enzyme binds substrate (e.g., nucleotide), but does not convert it to product (e.g., a DNA polymer).

[0220] k_{cat} (also called the turnover number of the enzyme) can be determined if the total enzyme concentration ($[E_T]$, i.e., the concentration of active sites) is known, since $V_{max}=k_{cat}[E_T]$. For situations in which the total enzyme concentration is difficult to measure, the ratio V_{max}/K_m is often used instead as a measure of efficiency. K_m and V_{max} can be determined, for example, from a Lineweaver-Burk plot of 1/V against 1/[S], where the y intercept represents $1/V_{max}$, the x intercept $-1/K_m$, and the slope K_m/V_{max}, or from an Eadie-Hofstee plot of V against V/[S], where the y intercept represents V_{max}, the x intercept V_{max}/K_m, and the slope $-K_m$. Software packages such as KinetAsyst™ or Enzfit (Biosoft, Cambridge, UK) can facilitate the determination of kinetic parameters from catalytic rate data.

[0221] For enzymes such as polymerases that have multiple substrates, varying the concentration of only one substrate while holding the others in suitable excess (e.g., effectively constant) concentration typically yields normal Michaelis-Menten kinetics.

[0222] In one embodiment, using pre-steady-state kinetics, the nucleotide concentration dependence of the rate k_{obs} (the observed first-order rate constant for dNTP incorporation) provides an estimate of the K_m for a ground state binding and the maximum rate of polymerization (k_{pol}). The k_{obs} is measured using a burst assay. The results of the assay are fitted with the Burst equation; Product = $A[1-\exp(-k_{obs}*t)]+k_{ss}*t$ where A represents amplitude an estimate of the concentration of the enzyme active site*s, k_{ss} is the observed steady-state rate constant and t is the reaction incubation time. The K_m for dNTP binding to the polymerase-DNA complex and the k_{pol} are calculated by fitting the dNTP concentration dependent change in the k_{obs} using the equation $k_{obs} = (k_{pol}*[S])*(K_m+[S])-1$ where [S] is the substrate concentration. Results are optionally obtained from a rapid-quench experiment (also called a quench-flow measurement), for example, based on the methods described in Johnson (1986) "Rapid kinetic analysis of mechanochemical adenosinetriphosphatases" Methods Enzymol.

134:677-705, Patel et al. (1991) "Pre-steady-state kinetic analysis of processive DNA replication including complete characterization of an exonuclease-deficient mutant" Biochemistry 30(2):511-25, and Tsai and Johnson (2006) "A new paradigm for DNA polymerase specificity" Biochemistry 45(32):9675-87.

[0223] Parameters such as rate of binding of a nucleotide analog or template by the recombinant polymerase, rate of product release by the recombinant polymerase, or branching rate of the recombinant polymerase can also be determined, and optionally compared to that of the first polymerase (e.g., a corresponding wild-type polymerase).

[0224] For a more thorough discussion of enzyme kinetics, see, e.g., Berg, Tymoczko, and Stryer (2002) Biochemistry, Fifth Edition, W. H. Freeman; Creighton (1984) Proteins: Structures and Molecular Principles, W. H. Freeman; and Fersht (1985) Enzyme Structure and Mechanism, Second Edition, W. H. Freeman.

[0225] In one aspect, the improved property of the enzymes of the invention is measured with reference to a model analog or analog set and compared with a given parental enzyme. For example, in the case of enzymes derived from a Φ29 parental enzyme, where the improvement being sought is a decrease in stability of the closed complex, an improved enzyme of the invention (i.e., an enzyme with an increased branching fraction) would have a higher k_{off} than the parental enzyme, e.g., wild type Φ29. While the foregoing may be used as a characterization tool, it in no way is intended as a specifically limiting reaction of the invention.

Screening Polymerases

[0226] Screening or other protocols can be used to determine whether a polymerase displays a modified activity for a nucleotide analog as compared to the first DNA polymerase. For example, k_{off}, k_{cat}, K_m, V_{max}, or k_{cat}/K_m of the recombinant DNA polymerase for the template or nucleotide or analog can be determined as discussed above.

[0227] In one desirable aspect, a library of recombinant DNA polymerases can be made and screened for these properties. For example, a plurality of members of the library can be made to include one or more mutations that increase branching fractions, delay translocation

and/or increase residence time, that are then screened for the properties of interest (e.g., increased branching fraction, delayed translocation or increased residence time). In general, the library can be screened to identify at least one member comprising a modified activity of interest.

[0228] Libraries of polymerases can be either physical or logical in nature. Moreover, any of a wide variety of library formats can be used. For example, polymerases can be fixed to solid surfaces in arrays of proteins. Similarly, liquid phase arrays of polymerases (e.g., in microwell plates) can be constructed for convenient high-throughput fluid manipulations of solutions comprising polymerases. Liquid, emulsion, or gel-phase libraries of cells that express recombinant polymerases can also be constructed, e.g., in microwell plates, or on agar plates. Phage display libraries of polymerases or polymerase domains (e.g., including the active site region or interdomain stability regions) can be produced. Likewise, yeast display libraries can be used. Instructions in making and using libraries can be found, e.g., in Sambrook, Ausubel and Berger, referenced herein.

[0229] For the generation of libraries involving fluid transfer to or from microtiter plates, a fluid handling station is optionally used. Several "off the shelf" fluid handling stations for performing such transfers are commercially available, including e.g., the Zymate systems from Caliper Life Sciences (Hopkinton, MA) and other stations which utilize automatic pipettors, e.g., in conjunction with the robotics for plate movement (e.g., the ORCA® robot, which is used in a variety of laboratory systems available, e.g., from Beckman Coulter, Inc. (Fullerton, CA).

[0230] In an alternate embodiment, fluid handling is performed in microchips, e.g., involving transfer of materials from microwell plates or other wells through microchannels on the chips to destination sites (microchannel regions, wells, chambers or the like). Commercially available microfluidic systems include those from Hewlett-Packard/Agilent Technologies (e.g., the HP2100 bioanalyzer) and the Caliper High Throughput Screening System. The Caliper High Throughput Screening System provides one example interface between standard microwell library formats and Labchip technologies. RainDance Technologies' nanodroplet platform provides another method for handling large numbers of

spatially separated reactions. Furthermore, the patent and technical literature includes many examples of microfluidic systems which can interface directly with microwell plates for fluid handling.

Desirable Properties

[0231] The polymerases of the invention can include any of a variety of modified properties towards natural or nucleotide analogs or analogs, depending on the application, including increased branching fractions, delayed translocation, increased nucleotide or nucleotide analog residence time, greater processivity, etc. For example, branching rates can be directly monitored in high-throughput SMS reactions using known templates. Branching or translocation rates can be screened for or against in selecting a polymerase of the invention, e.g., by screening enzymes based on kinetic or product formation properties. Nucleotide analog residence time is readily determined by observing signal pulse widths as detected by a signal detector of a sequencing system.

[0232] For example, improvements in a dissociation rate (or improved processivity) of 30% or more, e.g., about 50%, 75%, or even 100% or more can be screened for in identifying polymerases that display decreased translocation rates. Similarly, detecting mutant polymerases that exhibit branch rates of more than 30%, e.g., 40% or more, preferably 50% or more, or even 75% or more is a feature of the invention.

AFFINITY TAGS AND OTHER OPTIONAL POLYMERASE FEATURES

[0233] The recombinant DNA polymerases optionally include additional features exogenous or heterologous to the polymerases. For example, the recombinant polymerases optionally include one or more exogenous affinity tags, e.g., purification or substrate binding tags, such as a GST tag, an HA tag sequence, a plurality of GST tags, a plurality of HA tag sequences, a SNAP-tag, a c-myc tag, a c-myc fusion, or the like. These and other features useful in the context of binding a polymerase to a surface are optionally included, e.g., to orient and/or protect the polymerase active site when the polymerase is bound to a surface. Other useful features include recombinant dimer domains of the enzyme, and, e.g., large extraneous polypeptide domains coupled to the polymerase distal to the active site. For example, for Φ29, the active site is in the C terminal region of the protein, and added surface binding elements

(extra domains, GST tags, etc.) are typically located in the N-terminal region to avoid interfering with the active site when the polymerase is coupled to a surface.

[0234] In general, surface binding elements and purification tags that can be added to the polymerase (recombinantly or, e.g., chemically) include, e.g., biotin, avidin, GST sequences, modified GST sequences, e.g., that are less likely to form dimers, BiTag sequences, S tags, SNAP-tags, enterokinase sites, thrombin sites, antibodies or antibody domains, antibody fragments, antigens, receptors, receptor domains, receptor fragments, ligands, dyes, acceptors, quenchers, or combinations thereof.

[0235] Multiple surface binding domains can be added to orient the polypeptide relative to a surface and/or to increase binding of the polymerase to the surface. By binding a surface at two or more sites, through two or more separate tags, the polymerase is held in a relatively fixed orientation with respect to the surface. Additional details on fixing a polymerase to a surface, attaching tags, and the like are found in WO 2007/075987 ACTIVE SURFACE COUPLED POLYMERASES by Hanzel et al., and WO 2007/076057 PROTEIN ENGINEERING STRATEGIES TO OPTIMIZE ACTIVITY OF SURFACE ATTACHED PROTEINS by Hanzel et al.

MAKING AND ISOLATING RECOMBINANT POLYMERASES

[0236] Generally, nucleic acids encoding a polymerase of the invention can be made by cloning, recombination, in vitro synthesis, in vitro amplification and/or other available methods. A variety of recombinant methods can be used for expressing an expression vector that encodes a polymerase of the invention, e.g., a modified polymerase with an increased branching rate, delayed translocation or increased nucleotide analog residence time. Recombinant methods for making nucleic acids, expression and isolation of expressed products are well known and described in the art. For example, when modifying the active site to increase branching properties, features are selected (e.g., by modeling, though random approaches can also be used) that hinder steric access of the nucleotide analog to the active site, and/or that interfere with charge-charge or hydrophobic interactions between a given nucleotide analog and the polymerase target. Methods for making and selecting mutations in the active site of polymerases, including for modifying steric features in or near the active site to permit

improved access by nucleotide analogs are found, e.g., in WO 2007/076057 POLYMERASES FOR NUCLEOTIDE ANALOG INCORPORATION by Hanzel et al., and PCT/US2007/022459 POLYMERASE ENZYMES AND REAGENTS FOR ENHANCED NUCLEIC ACID SEQUENCING by Rank et al.

[0237] Additional useful references for mutation, recombinant and in vitro nucleic acid manipulation methods (including cloning, expression, PCR, and the like) include Berger and Kimmel, Guide to Molecular Cloning Techniques, Methods in Enzymology volume 152 Academic Press, Inc., San Diego, CA (Berger); Kaufman et al. (2003) Handbook of Molecular and Cellular Methods in Biology and Medicine Second Edition Ceske (ed) CRC Press (Kaufman); and The Nucleic Acid Protocols Handbook Ralph Rapley (ed) (2000) Cold Spring Harbor, Humana Press Inc (Rapley); Chen et al. (ed) PCR Cloning Protocols, Second Edition (Methods in Molecular Biology, volume 192) Humana Press; and in Viljoen et al. (2005) Molecular Diagnostic PCR Handbook Springer, ISBN 1402034032.

[0238] In addition, a plethora of kits are commercially available for the purification of plasmids or other relevant nucleic acids from cells, (*see*, e.g., EasyPrep™, FlexiPrep™, both from Pharmacia Biotech; StrataClean™, from Stratagene; and, QIAprep™ from Qiagen). Any isolated and/or purified nucleic acid can be further manipulated to produce other nucleic acids, used to transfect cells, incorporated into related vectors to infect organisms for expression, and/or the like. Typical cloning vectors contain transcription and translation terminators, transcription and translation initiation sequences, and promoters useful for regulation of the expression of the particular target nucleic acid. The vectors optionally comprise generic expression cassettes containing at least one independent terminator sequence, sequences permitting replication of the cassette in eukaryotes, or prokaryotes, or both, (e.g., shuttle vectors) and selection markers for both prokaryotic and eukaryotic systems. Vectors are suitable for replication and integration in prokaryotes, eukaryotes, or both.

[0239] Other useful references, e.g. for cell isolation and culture (e.g., for subsequent nucleic acid isolation) include Freshney (1994) Culture of Animal Cells, a Manual of Basic Technique, third edition, Wiley- Liss, New York and the references cited therein; Payne *et al.* (1992) Plant Cell and Tissue Culture in Liquid Systems John Wiley & Sons, Inc. New York,

NY; Gamborg and Phillips (eds) (1995) Plant Cell, Tissue and Organ Culture; Fundamental Methods Springer Lab Manual, Springer-Verlag (Berlin Heidelberg New York) and Atlas and Parks (eds) The Handbook of Microbiological Media (1993) CRC Press, Boca Raton, FL.

[0240] A variety of protein isolation and detection methods are known and can be used to isolate polymerases, e.g., from recombinant cultures of cells expressing the recombinant polymerases of the invention. A variety of protein isolation and detection methods are well known in the art, including, e.g., those set forth in R. Scopes, Protein Purification, Springer-Verlag, N.Y. (1982); Deutscher, Methods in Enzymology Vol. 182: Guide to Protein Purification, Academic Press, Inc. N.Y. (1990); Sandana (1997) Bioseparation of Proteins, Academic Press, Inc.; Bollag et al. (1996) Protein Methods, 2nd Edition Wiley-Liss, NY; Walker (1996) The Protein Protocols Handbook Humana Press, NJ, Harris and Angal (1990) Protein Purification Applications: A Practical Approach IRL Press at Oxford, Oxford, England; Harris and Angal Protein Purification Methods: A Practical Approach IRL Press at Oxford, Oxford, England; Scopes (1993) Protein Purification: Principles and Practice 3rd Edition Springer Verlag, NY; Janson and Ryden (1998) Protein Purification: Principles, High Resolution Methods and Applications, Second Edition Wiley-VCH, NY; and Walker (1998) Protein Protocols on CD-ROM Humana Press, NJ; and the references cited therein. Additional details regarding protein purification and detection methods can be found in Satinder Ahuja ed., Handbook of Bioseparations, Academic Press (2000).

NUCLEIC ACID AND POLYPEPTIDE SEQUENCE AND VARIANTS

[0241] As described herein, the invention provides polynucleotide sequences encoding, e.g., a polymerase as described herein. Examples of polymerase sequences that include features are found herein, e.g., increased branching fractions as in Table A. However, one of skill in the art will immediately appreciate that the invention is not limited to the specifically exemplified sequences. For example, one of skill will appreciate that the invention also provides, e.g., many related sequences with the functions described herein, e.g., polynucleotides and polypeptides encoding conservative variants of a polymerase of Table A or any other specifically listed polymerase herein. Combinations of any of the mutations noted herein or combinations of any of the mutations herein in combination with those noted in other

available references relating to improved polymerases, such as Hanzel et WO 2007/076057 POLYMERASES FOR NUCLEOTIDE ANALOG INCORPORATION; Rank et al. PCT/US2007/022459 POLYMERASE ENZYMES AND REAGENTS FOR ENHANCED NUCLEIC ACID SEQUENCING; Hanzel et al. WO 2007/075987 ACTIVE SURFACE COUPLED POLYMERASES; and Hanzel et al. WO 2007/076057 PROTEIN ENGINEERING STRATEGIES TO OPTIMIZE ACTIVITY OF SURFACE ATTACHED PROTEINS are also features of the invention

[0242] Accordingly, the invention provides a variety of polypeptides (polymerases) and polynucleotides (nucleic acids that encode polymerases). Example polynucleotides of the invention include, e.g., a polynucleotide encoding a polymerase as set forth in Table A or a polynucleotide that is complementary to or that encodes a polynucleotide sequence thereof (e.g., where the given sequence is a DNA, an RNA is one example of a sequence that encodes the DNA, e.g., via reverse transcription). A polynucleotide of the invention also optionally includes any polynucleotide that encodes a polymerase of Table A. Because of the degeneracy of the genetic code, many polynucleotides equivalently encode a given polymerase sequence. Similarly, an artificial or recombinant nucleic acid that hybridizes to a polynucleotide indicated above under highly stringent conditions over substantially the entire length of the nucleic acid (and is other than a naturally occurring polynucleotide) is a polynucleotide of the invention. In one embodiment, a composition includes a polypeptide of the invention and an excipient (e.g., buffer, water, pharmaceutically acceptable excipient, etc.). The invention also provides an antibody or antisera specifically immunoreactive with a polypeptide of the invention (e.g., that specifically recognizes a feature of the polymerase that confers increased branching, delayed translocation or increased nucleotide analog residence time.

[0243] In certain embodiments, a vector (e.g., a plasmid, a cosmid, a phage, a virus, etc.) comprises a polynucleotide of the invention. In one embodiment, the vector is an expression vector. In another embodiment, the expression vector includes a promoter operably linked to one or more of the polynucleotides of the invention. In another embodiment, a cell comprises a vector that includes a polynucleotide of the invention.

[0244] One of skill will also appreciate that many variants of the disclosed sequences are included in the invention. For example, conservative variations of the disclosed sequences that yield a functionally similar sequence are included in the invention. Variants of the nucleic acid polynucleotide sequences, wherein the variants hybridize to at least one disclosed sequence, are considered to be included in the invention. Unique subsequences of the sequences disclosed herein, as determined by, e.g., standard sequence comparison techniques, are also included in the invention.

Conservative Variations

[0245] Owing to the degeneracy of the genetic code, "silent substitutions" (*i.e.*, substitutions in a nucleic acid sequence which do not result in an alteration in an encoded polypeptide) are an implied feature of *every* nucleic acid sequence that encodes an amino acid sequence. Similarly, "conservative amino acid substitutions," where one or a limited number of amino acids in an amino acid sequence are substituted with different amino acids with highly similar properties, are also readily identified as being highly similar to a disclosed construct. Such conservative variations of each disclosed sequence are a feature of the present invention.

[0246] "Conservative variations" of a particular nucleic acid sequence refers to those nucleic acids which encode identical or essentially identical amino acid sequences, or, where the nucleic acid does not encode an amino acid sequence, to essentially identical sequences. One of skill will recognize that individual substitutions, deletions or additions which alter, add or delete a single amino acid or a small percentage of amino acids (typically less than 5%, more typically less than 4%, 2% or 1%) in an encoded sequence are "conservatively modified variations" where the alterations result in the deletion of an amino acid, addition of an amino acid, or substitution of an amino acid with a chemically similar amino acid, while retaining the relevant mutational feature (for example, the conservative substitution can be of a residue distal to the active site region, or distal to an interdomain stability region). Thus, "conservative variations" of a listed polypeptide sequence of the present invention include substitutions of a small percentage, typically less than 5%, more typically less than 2% or 1%, of the amino acids of the polypeptide sequence, with an amino acid of the same conservative substitution group. Finally, the addition of sequences which do not alter the encoded activity of a nucleic acid

molecule, such as the addition of a non-functional or tagging sequence (introns in the nucleic acid, poly His or similar sequences in the encoded polypeptide, etc.), is a conservative variation of the basic nucleic acid or polypeptide.

[0247] Conservative substitution tables providing functionally similar amino acids are well known in the art, where one amino acid residue is substituted for another amino acid residue having similar chemical properties (e.g., aromatic side chains or positively charged side chains), and therefore does not substantially change the functional properties of the polypeptide molecule. The following sets forth example groups that contain natural amino acids of like chemical properties, where substitutions within a group is a "conservative substitution".

TABLE C
Conservative Amino Acid Substitutions

Nonpolar and/or Aliphatic Side Chains	Polar, Uncharged Side Chains	Aromatic Side Chains	Positively Charged Side Chains	Negatively Charged Side Chains
Glycine	Serine			
Alanine	Threonine	Phenylalanine	Lysine	Aspartate
Valine	Cysteine	Tyrosine	Arginine	Glutamate
Leucine	Methionine	Tryptophan	Histidine	
Isoleucine	Asparagine			
Proline	Glutamine			

Nucleic Acid Hybridization

[0248] Comparative hybridization can be used to identify nucleic acids of the invention, including conservative variations of nucleic acids of the invention. In addition, target nucleic acids which hybridize to a nucleic acid encoding a polymerase of Table A or any other specifically listed polymerase herein under high, ultra-high and ultra-ultra high stringency conditions, where the nucleic acids encode mutations corresponding to those noted in Table A or other listed polymerases are a feature of the invention. Examples of such nucleic acids include those with one or a few silent or conservative nucleic acid substitutions as compared to a given nucleic acid sequence encoding a polymerase of Table A (or other exemplified

polymerase), where any conservative substitutions are for residues other than those noted in Table A or elsewhere as being relevant to a feature of interest (improved complex stability, decreased branch rate formation, etc.).

[0249] A test nucleic acid is said to specifically hybridize to a probe nucleic acid when it hybridizes at least 50% as well to the probe as to the perfectly matched complementary target, i.e., with a signal to noise ratio at least half as high as hybridization of the probe to the target under conditions in which the perfectly matched probe binds to the perfectly matched complementary target with a signal to noise ratio that is at least about 5x-10x as high as that observed for hybridization to any of the unmatched target nucleic acids.

[0250] Nucleic acids "hybridize" when they associate, typically in solution. Nucleic acids hybridize due to a variety of well characterized physico-chemical forces, such as hydrogen bonding, solvent exclusion, base stacking and the like. An extensive guide to the hybridization of nucleic acids is found in Tijssen (1993) *Laboratory Techniques in Biochemistry and Molecular Biology--Hybridization with Nucleic Acid Probes* part I chapter 2, "Overview of principles of hybridization and the strategy of nucleic acid probe assays," (Elsevier, New York), as well as in Current Protocols in Molecular Biology, Ausubel et al., eds., Current Protocols, a joint venture between Greene Publishing Associates, Inc. and John Wiley & Sons, Inc., (supplemented through 2004) ("Ausubel"); Hames and Higgins (1995) Gene Probes 1 IRL Press at Oxford University Press, Oxford, England, (Hames and Higgins 1) and Hames and Higgins (1995) Gene Probes 2 IRL Press at Oxford University Press, Oxford, England (Hames and Higgins 2) provide details on the synthesis, labeling, detection and quantification of DNA and RNA, including oligonucleotides.

[0251] An example of stringent hybridization conditions for hybridization of complementary nucleic acids which have more than 100 complementary residues on a filter in a Southern or northern blot is 50% formalin with 1 mg of heparin at 42°C, with the hybridization being carried out overnight. An example of stringent wash conditions is a 0.2x SSC wash at 65°C for 15 minutes (*see,* Sambrook, *supra* for a description of SSC buffer). Often the high stringency wash is preceded by a low stringency wash to remove background probe signal. An example low stringency wash is 2x SSC at 40°C for 15 minutes. In general, a signal to noise

ratio of 5x (or higher) than that observed for an unrelated probe in the particular hybridization assay indicates detection of a specific hybridization.

[0252] "Stringent hybridization wash conditions" in the context of nucleic acid hybridization experiments such as Southern and northern hybridizations are sequence dependent, and are different under different environmental parameters. An extensive guide to the hybridization of nucleic acids is found in Tijssen (1993), *supra.* and in Hames and Higgins, 1 and 2. Stringent hybridization and wash conditions can easily be determined empirically for any test nucleic acid. For example, in determining stringent hybridization and wash conditions, the hybridization and wash conditions are gradually increased (e.g., by increasing temperature, decreasing salt concentration, increasing detergent concentration and/or increasing the concentration of organic solvents such as formalin in the hybridization or wash), until a selected set of criteria are met. For example, in highly stringent hybridization and wash conditions, the hybridization and wash conditions are gradually increased until a probe binds to a perfectly matched complementary target with a signal to noise ratio that is at least 5x as high as that observed for hybridization of the probe to an unmatched target.

[0253] "Very stringent" conditions are selected to be equal to the thermal melting point (T_m) for a particular probe. The T_m is the temperature (under defined ionic strength and pH) at which 50% of the test sequence hybridizes to a perfectly matched probe. For the purposes of the present invention, generally, "highly stringent" hybridization and wash conditions are selected to be about 5° C lower than the T_m for the specific sequence at a defined ionic strength and pH.

[0254] "Ultra high-stringency" hybridization and wash conditions are those in which the stringency of hybridization and wash conditions are increased until the signal to noise ratio for binding of the probe to the perfectly matched complementary target nucleic acid is at least 10x as high as that observed for hybridization to any of the unmatched target nucleic acids. A target nucleic acid which hybridizes to a probe under such conditions, with a signal to noise ratio of at least ½ that of the perfectly matched complementary target nucleic acid is said to bind to the probe under ultra-high stringency conditions.

[0255] Similarly, even higher levels of stringency can be determined by gradually increasing the hybridization and/or wash conditions of the relevant hybridization assay. For example, those in which the stringency of hybridization and wash conditions are increased until the signal to noise ratio for binding of the probe to the perfectly matched complementary target nucleic acid is at least 10x, 20X, 50X, 100X, or 500X or more as high as that observed for hybridization to any of the unmatched target nucleic acids. A target nucleic acid which hybridizes to a probe under such conditions, with a signal to noise ratio of at least ½ that of the perfectly matched complementary target nucleic acid is said to bind to the probe under ultra-ultra-high stringency conditions.

[0256] Nucleic acids that do not hybridize to each other under stringent conditions are still substantially identical if the polypeptides which they encode are substantially identical. This occurs, *e.g.*, when a copy of a nucleic acid is created using the maximum codon degeneracy permitted by the genetic code.

Unique Subsequence

[0257] In some aspects, the invention provides a nucleic acid that comprises a unique subsequence in a nucleic acid that encodes a polymerase of Table A. The unique subsequence may be unique as compared to a nucleic acid corresponding to, e.g., a wild type Φ29. Alignment can be performed using, e.g., BLAST set to default parameters. Any unique subsequence is useful, e.g., as a probe to identify the nucleic acids of the invention.

[0258] Similarly, the invention includes a polypeptide which comprises a unique subsequence in a polymerase of Table A. Here, the unique subsequence is unique as compared to, e.g., wild type Φ29 or previously characterized mutation thereof.

[0259] The invention also provides for target nucleic acids which hybridize under stringent conditions to a unique coding oligonucleotide which encodes a unique subsequence in a polypeptide selected from the sequences of Table A, wherein the unique subsequence is unique as compared to a polypeptide corresponding to wild type Φ29. Unique sequences are determined as noted above.

Sequence comparison, identity and homology

[0260] The terms "identical" or "percent identity," in the context of two or more nucleic acid or polypeptide sequences, refer to two or more sequences or subsequences that are the same or have a specified percentage of amino acid residues or nucleotides that are the same, when compared and aligned for maximum correspondence, as measured using one of the sequence comparison algorithms described below (or other algorithms available to persons of skill) or by visual inspection.

[0261] The phrase "substantially identical," in the context of two nucleic acids or polypeptides (*e.g.*, DNAs encoding a polymerase, or the amino acid sequence of a polymerase) refers to two or more sequences or subsequences that have at least about 60%, about 80%, about 90-95%, about 98%, about 99% or more nucleotide or amino acid residue identity, when compared and aligned for maximum correspondence, as measured using a sequence comparison algorithm or by visual inspection. Such "substantially identical" sequences are typically considered to be "homologous," without reference to actual ancestry. Preferably, the "substantial identity" exists over a region of the sequences that is at least about 50 residues in length, more preferably over a region of at least about 100 residues, and most preferably, the sequences are substantially identical over at least about 150 residues, or over the full length of the two sequences to be compared.

[0262] Proteins and/or protein sequences are "homologous" when they are derived, naturally or artificially, from a common ancestral protein or protein sequence. Similarly, nucleic acids and/or nucleic acid sequences are homologous when they are derived, naturally or artificially, from a common ancestral nucleic acid or nucleic acid sequence. Homology is generally inferred from sequence similarity between two or more nucleic acids or proteins (or sequences thereof). The precise percentage of similarity between sequences that is useful in establishing homology varies with the nucleic acid and protein at issue, but as little as 25% sequence similarity over 50, 100, 150 or more residues is routinely used to establish homology. Higher levels of sequence similarity, e.g., 30%, 40%, 50%, 60%, 70%, 80%, 90%, 95%, or 99% or more, can also be used to establish homology. Methods for determining sequence similarity percentages (e.g., BLASTP and BLASTN using default parameters) are described herein and are generally available.

[0263] For sequence comparison and homology determination, typically one sequence acts as a reference sequence to which test sequences are compared. When using a sequence comparison algorithm, test and reference sequences are input into a computer, subsequence coordinates are designated, if necessary, and sequence algorithm program parameters are designated. The sequence comparison algorithm then calculates the percent sequence identity for the test sequence(s) relative to the reference sequence, based on the designated program parameters.

[0264] Optimal alignment of sequences for comparison can be conducted, *e.g.*, by the local homology algorithm of Smith & Waterman, Adv. Appl. Math. 2:482 (1981), by the homology alignment algorithm of Needleman & Wunsch, J. Mol. Biol. 48:443 (1970), by the search for similarity method of Pearson & Lipman, Proc. Nat'l. Acad. Sci. USA 85:2444 (1988), by computerized implementations of these algorithms (GAP, BESTFIT, FASTA, and TFASTA in the Wisconsin Genetics Software Package, Genetics Computer Group, 575 Science Dr., Madison, WI), or by visual inspection (*see generally* Current Protocols in Molecular Biology, Ausubel et al., eds., Current Protocols, a joint venture between Greene Publishing Associates, Inc. and John Wiley & Sons, Inc., supplemented through 2004).

[0265] One example of an algorithm that is suitable for determining percent sequence identity and sequence similarity is the BLAST algorithm, which is described in Altschul *et al.*, J. Mol. Biol. 215:403-410 (1990). Software for performing BLAST analyses is publicly available through the National Center for Biotechnology Information. This algorithm involves first identifying high scoring sequence pairs (HSPs) by identifying short words of length W in the query sequence, which either match or satisfy some positive-valued threshold score T when aligned with a word of the same length in a database sequence. T is referred to as the neighborhood word score threshold (Altschul *et al.*, *supra*). These initial neighborhood word hits act as seeds for initiating searches to find longer HSPs containing them. The word hits are then extended in both directions along each sequence for as far as the cumulative alignment score can be increased. Cumulative scores are calculated using, for nucleotide sequences, the parameters M (reward score for a pair of matching residues; always > 0) and N (penalty score for mismatching residues; always < 0). For amino acid sequences, a scoring matrix is used to calculate the cumulative score. Extension of the word hits in each direction are halted when:

the cumulative alignment score falls off by the quantity X from its maximum achieved value; the cumulative score goes to zero or below, due to the accumulation of one or more negative-scoring residue alignments; or the end of either sequence is reached. The BLAST algorithm parameters W, T, and X determine the sensitivity and speed of the alignment. The BLASTN program (for nucleotide sequences) uses as defaults a wordlength (W) of 11, an expectation (E) of 10, a cutoff of 100, M=5, N=-4, and a comparison of both strands. For amino acid sequences, the BLASTP program uses as defaults a wordlength (W) of 3, an expectation (E) of 10, and the BLOSUM62 scoring matrix (*see* Henikoff & Henikoff (1989) Proc. Natl. Acad. Sci. USA 89:10915).

[0266] In addition to calculating percent sequence identity, the BLAST algorithm also performs a statistical analysis of the similarity between two sequences (*see, e.g.*, Karlin & Altschul, Proc. Nat'l. Acad. Sci. USA 90:5873-5787 (1993)). One measure of similarity provided by the BLAST algorithm is the smallest sum probability (P(N)), which provides an indication of the probability by which a match between two nucleotide or amino acid sequences would occur by chance. For example, a nucleic acid is considered similar to a reference sequence if the smallest sum probability in a comparison of the test nucleic acid to the reference nucleic acid is less than about 0.1, more preferably less than about 0.01, and most preferably less than about 0.001.

[0267] While the foregoing invention has been described in some detail for purposes of clarity and understanding, it will be clear to one skilled in the art from a reading of this disclosure that various changes in form and detail can be made without departing from the true scope of the invention. For example, all the techniques and apparatus described above can be used in various combinations. All publications, patents, patent applications, and/or other documents cited in this application are incorporated by reference in their entirety for all purposes to the same extent as if each individual publication, patent, patent application, and/or other document were individually indicated to be incorporated by reference for all purposes.

CLAIMS

WHAT IS CLAIMED IS:

1. A method of determining which nucleotide is incorporated at a site during a template dependent polymerization reaction, the method comprising:

incorporating the nucleotide into a nucleic acid polymer whereby signal pulses or signatures are generated from branch fraction nonincorporation events and, optionally, actual nucleotide incorporation events, for the site;

monitoring a time course of at least branch fraction signal pulses or signatures produced by the polymerization reaction; and,

assigning which nucleotide is incorporated at the site, using at least signal pulses or signatures from branch fraction nonincorporation sampling events.

2. The method of claim **1**, wherein the method comprises counting or estimating the number of redundant iterative sampling signal pulses per incorporation event, or determining an average number of redundant signal pulses per incorporation event.

3. The method of claim **1**, wherein the polymerization reaction is a high branch fraction polymerization reaction.

4. The method of claim **3**, wherein the branch fraction is at least about 80% or more.

5. The method of claim **1**, wherein the reaction comprises at least one species of metal ion, which metal ion increases the frequency of branch fraction nonincorporation events in the reaction.

6. The method of claim **5**, wherein the at least one species of metal ion is selected from Mg^{++}, Mn^{++}, Zn^{++}, Co^{++}, Ca^{++}, Fe^{++}, Cr^{++} and Sr^{++}.

7. The method of claim **5**, wherein the reaction comprises Mg^{++} and Mn^{++}.

8. The method of claim **7**, wherein the concentration of Mg^{++} is higher than the concentration of Mn^{++}.

9. The method of claim **1**, wherein the reaction comprises a Φ29, B103, GA-1, PZA, Φ15, BS32, M2Y, Nf, G1, Cp-1, PRD1, PZE, SF5, Cp-5, Cp-7, PR4, PR5, PR722, L17, T4 or T7 DNA polymerase, or a modified recombinant DNA polymerase thereof.

10. The method of claim **9**, wherein the modified recombinant polymerase exhibits a higher branching fraction as compared to a corresponding wild-type polymerase.

11. The method of claim **9**, wherein the modified recombinant polymerase comprises an exonuclease rate that is about 10% to 50% as compared to its polymerization rate.

12. The method of claim **1**, wherein the branch fraction nonincorporation events comprise iterative sampling of labeled unincorporatable nucleotide analogs.

13. The method of claim **12**, wherein actual nucleotide incorporation events comprise incorporation of unlabeled nucleotides.

14. The method of claim **12**, wherein actual nucleotide incorporation events comprise incorporation of nucleotides that are differentially labeled as compared to the unincorporatable nucleotides.

15. The method of claim **12**, wherein the labeled unincorporatable nucleotide analogs comprise a link between an alpha and beta phosphate group that is not hydrolyzable by a polymerase enzyme.

16. The method of claim **1**, wherein a polymerase enzyme, polymerase reaction conditions, and/or polymerase reaction substrates are selected such that the polymerization reaction exhibits two kinetically observable steps within an observable phase of the polymerase reaction.

17. The method of claim **16**, wherein the two kinetically observable steps are each steps which proceed in a bright phase.

18. The method of claim **16**, wherein the two kinetically observable steps are each steps which proceed in a dark phase.

19. The method of claim **16**, wherein the polymerase enzyme comprises a modified recombinant Φ29, B103, GA-1, PZA, Φ15, BS32, M2Y, Nf, G1, Cp-1, PRD1, PZE, SF5, Cp-5, Cp-7, PR4, PR5, PR722, L17, T4 or T7 DNA polymerase.

20. The method of claim **16**, wherein the polymerase reaction conditions comprise one or more of: a selected metal cofactor concentration, a selected pH, a selected temperature, an enzyme activity modulator, D_2O, an organic solvent, and a buffer.

21. The method of claim **1**, wherein the branch fraction nonincorporation events comprise noncognate branch fraction nonincorporation events.

22. The method of claim **21**, wherein the branch fraction signal pulses or signatures are generated from noncognate branching events of a nucleotide analog comprising a base selected from: guanine and thymine.

23. The method of claim **1**, wherein the reaction is reacted in a DNA sequencing system.

24. The method of claim **23**, wherein the DNA sequencing system comprises a zero mode waveguide or nanohole.

25. The method of claim **1**, wherein assigning the nucleotide comprises applying a statistical model to the signal pulses or signatures generated from branch fraction nonincorporation events, signal pulses generated from actual nucleotide incorporation events, or both, which statistical model assigns a likelihood that the signal pulses or signatures correspond to an incorporation event.

26. The method of claim **1**, further comprising:

performing an additional template dependent polymerization reaction under high processivity reaction conditions;

monitoring a second time course of signal pulses or signatures produced by the additional polymerization reaction; and,

compiling sequencing information derived from the second time course of signal pulses or signatures with sequencing information derived from the time course of branch fraction signal pulses or signatures.

27. A multi-modal sequencing method, comprising:

performing a first template dependent sequencing reaction in a first mode comprising a first set of reaction conditions;

collecting initial sequencing information produced by the first sequencing reaction;

performing a second sequencing reaction of the template, or a copy thereof, in a second mode comprising a second set of reaction conditions;

collecting additional sequencing information produced by the second sequencing reaction; and,

compiling the initial and additional sequencing information to provide a sequence of at least a portion of the template.

28. The method of claim **27**, wherein the second sequencing reaction is produced by altering one or more reaction conditions of the first sequencing reaction, wherein the initial and additional sequencing information are collected in real time.

29. The method of claim **28**, wherein altering one or more reaction conditions comprises adding one or more polymerase cofactors to the first sequencing reaction.

30. The method of claim **28**, wherein altering one or more reaction conditions comprises adding Mn^{++} or Mg^{++} to the first sequencing reaction.

31. The method of claim **27**, wherein both the first and second sequencing reactions comprise single template molecule sequencing reactions, and wherein the first mode produces a higher branch fraction as compared to the second mode.

32. The method of claim **27**, wherein both the first and second sequencing reactions comprise single template molecule sequencing reactions, and wherein the second mode produces longer read lengths as compared to the first mode.

33. The method of claim **27**, wherein the template is a single circular template molecule, and the method comprises switching between the first and second modes.

34. A method of determining which of two or more labeled nucleotides is incorporated at a site during a template-dependent polymerization reaction, the method comprising:

incorporating the nucleotide into a nucleic acid polymer produced by the polymerization reaction, whereby signal pulses or signatures are generated;

monitoring the pulses or signatures; and,

using the presence of multiple pulses corresponding to the nucleotide, or identical molecules thereof, to assign which labeled nucleotide is incorporated at the site.

35. The method of claim **34**, wherein the multiple pulses comprise 2 to 20 pulses.

36. The method of claim **34**, wherein the multiple pulses are generated from branch fraction nonincorporation events.

37. The method of claim **36**, wherein the branch fraction nonincorporation events are induced by a sequencing composition that comprises at least one species of metal ion.

38. The method of claim **37**, wherein the metal ion is selected from: Mg^{++}, Mn^{++}, Zn^{++}, Co^{++}, Ca^{++}, Fe^{++}, Cr^{++} and Sr^{++}.

39. The method of claim **34**, wherein using the presence of multiple pulses comprises distinguishing incorporation and nonincorporation signals to assign which labeled nucleotide incorporated at the site.

40. A nucleic acid sequencing system that, during operation of the system, sequences a nucleic acid, the system comprising:

a signal detector, which detector detects at least signal pulses or signatures from branch fraction nonincorporation events during sequencing of a template nucleic acid;

system instructions or software that assign a sequence based upon detection of at least signal pulses or signatures from branch fraction nonincorporation events; and,

a user output module that displays the sequence to the user.

41. The nucleic acid sequencing system of claim **40**, wherein the signal pulses or signatures from branch fraction nonincorporation events are generated during a first sequencing mode, wherein the signal detector detects signal pulses generated during a low branch fraction second sequencing mode, and wherein the system instructions assign a sequence based upon detection of signal pulses or signatures from the first and second sequencing modes.

42. The nucleic acid sequencing system of claim **40**, wherein the system comprises a zero-mode waveguide or nanohole proximal to the signal detector, and wherein, during operation of the system, a sequencing reaction is contained by the zero-mode waveguide or nanohole.

43. A composition comprising a modified recombinant nucleic acid polymerase, which modified recombinant polymerase exhibits an altered property selected from:

a) an increased branching fraction during a polymerization by the polymerase;

b) an altered translocation property of the polymerase during a polymerization reaction; and,

c) a combination of any of (a)-(b),

wherein the altered property is altered as compared to a corresponding wild-type polymerase.

44. The composition of claim **43**, wherein the modified recombinant polymerase is a modified recombinant Φ29, B103, GA-1, PZA, Φ15, BS32, M2Y, Nf, G1, Cp-1, PRD1, PZE, SF5, Cp-5, Cp-7, PR4, PR5, PR722, L17, T4, or T7 polymerase.

45. The composition of claim **43**, wherein the modified recombinant polymerase that exhibits an increased branching fraction comprises at least one amino acid substitution or deletion or combination of substitutions or deletions selected from: N62D and Y454A; D362S; Y259H; F237Y; L381I; Y369H; H461Y; A377G; K138Q; H461D; A377S; N62D and K371Q; V118L; and K124R; wherein numbering of the residues is relative to a wild-type Φ29 polymerase of SEQ ID NO: 3.

46. The composition of claim **43**, wherein the modified recombinant polymerase that exhibits an increased branching fraction exhibits a branching fraction that is at least 50% greater than the branching fraction of a wild-type Φ29 polymerase of SEQ ID NO: 3.

47. The composition of claim **43**, wherein the modified recombinant polymerase that exhibits an increased branching fraction exhibits a branching fraction that is at least 100% greater than the branching fraction of a wild-type Φ29 polymerase of SEQ ID NO: 3.

48. The composition of claim **43**, wherein the modified recombinant polymerase that exhibits an increased branching fraction exhibits a branching fraction that is at least 200% greater than the branching fraction of a wild-type Φ29 polymerase of SEQ ID NO: 3.

49. The composition of claim **43**, wherein the modified recombinant DNA polymerase that exhibits an increased branching fraction exhibits increased exonuclease activity as compared to the corresponding wild type polymerase.

50. The composition of claim **49**, wherein the modified recombinant polymerase comprises an exonuclease rate that is about 10% to 50% as compared to its polymerization rate.

51. The composition of claim **43**, wherein the modified recombinant polymerase that exhibits an altered translocation property comprises a fusion protein, which fusion protein comprises at least a subsequence of the parental polymerase and at least one heterologous polypeptide sequence.

52. The composition of claim **51**, wherein the fusion protein comprises the at least one heterologous polypeptide sequence fused at or near the c-terminal end of the parental polymerase.

53. The composition of claim **52**, wherein the wild-type polymerase is a Φ29 polymerase.

54. The composition of claim **51**, wherein the heterologous polypeptide sequence comprises at least one charged amino acid.

55. The composition of claim **54**, wherein the at least one charged amino acid is histidine.

56. The composition of claim **51**, wherein the fusion protein comprises a linker between the at least a subsequence of the parental polymerase and the heterologous polypeptide sequence.

57. The composition of claim **56**, wherein the linker comprises a Ser3Gly linker.

58. The composition of claim **43**, wherein the modified recombinant polymerase that exhibits an altered translocation property comprises at least one amino acid substitution or deletion or combination of substitutions or deletions selected from: Asp570Lys; Asp570Ala; Asn313Lys; Asn313Ala; Gln303Lys; Gln303Ala; Gly532Ser; Met533delet; Cys530delet; Met533delet and Cys530delet; Gly532delet; Ala531Gly; Thr573Lys; Thr573Ala; Asn396Lys;

Thr571Lys; Thr571Ala; Thr534Lys; Thr534Ala; Asp535Lys; Asp535Ala; Lys529Ala; and Lys529Asn; wherein numbering of the residue positions is relative to a wild-type Φ29 polymerase of SEQ ID NO: 3.

59. The composition of claim **43**, wherein the modified recombinant DNA polymerase that exhibits an altered translocation property exhibits a delay in translocation as compared to the corresponding wild-type polymerase.

60. The composition of claim **59**, wherein the delay in translocation is at least about 2.5 times greater than that of the corresponding wild-type polymerase.

61. The composition of claim **59**, wherein the delay in translocation is at least about 10 times greater than that of the corresponding wild-type polymerase.

62. The composition of claim **59**, wherein the delay in translocation is at least about 15 times greater than that of the corresponding wild-type polymerase.

63. The composition of claim **43**, wherein the modified recombinant polymerase exhibits an increased nucleotide or nucleotide analog residence time as compared to the corresponding wild-type polymerase.

64. The composition of claim **43**, wherein the modified recombinant polymerase exhibits increased processivity as compared to the corresponding wild-type polymerase.

65. The composition of claim **43**, further comprising a phosphate-labeled nucleotide analog and a DNA template, wherein the modified recombinant polymerase incorporates the nucleotide analog into a copy nucleic acid in response to the DNA template.

66. The composition of claim **43**, wherein the composition is present in a DNA sequencing system.

67. The composition of claim **66**, wherein the DNA sequencing system comprises at least one zero mode waveguide or nanohole.

68. The composition of claim **43**, wherein the modified recombinant polymerase is immobilized on a surface.

69. A method of sequencing a nucleic acid template, the method comprising:

a) providing a reaction mixture comprising:

i) the nucleic acid template;

ii) a replication initiating moiety that complexes with or is integral to the template;

iii) the modified recombinant nucleic acid polymerase of claim **43**, wherein the polymerase is capable of replicating at least a portion of the template using the moiety in a template-dependent polymerization reaction; and

iv) one or more nucleotides and/or nucleotide analogs;

b) subjecting the reaction mixture to a polymerization reaction in which the modified recombinant polymerase replicates at least a portion of the template in a template-dependent manner, whereby the one or more nucleotides and/or nucleotide analogs are incorporated into the resulting copy nucleic acid; and

c) identifying a time sequence of incorporation of the one or more nucleotides and/or nucleotide analogs into the resulting copy nucleic acid.

70. The method of claim **69**, wherein the modified recombinant polymerase exhibits increased processivity relative to the wild-type polymerase.

71. The method of claim **69**, wherein identifying the time sequence of incorporation comprises observing more than one signal pulse per nucleotide incorporation event.

72. The method of claim **69**, wherein the polymerization reaction is performed in a zero mode waveguide or nanohole.

73. A method of making a nucleic acid, the method comprising:

a) providing a reaction mixture comprising:

i) a template,

ii) a replication initiating moiety that complexes with or is integral to the template,

iii) the modified recombinant nucleic acid polymerase of claim **43**, wherein the polymerase is capable of replicating at least a portion of the template using the moiety in a template-dependent polymerase reaction, and

iv) one or more nucleotides and/or nucleotide analogs; and

b) reacting the mixture such that the polymerase replicates at least a portion of the template in a template-dependent manner, whereby the one or more nucleotides and/or nucleotide analogs are incorporated into the resulting nucleic acid.

74. The method of claim **73**, the method comprising detecting incorporation of at least one of the nucleotides and/or nucleotide analogs.

75. The method of claim **74**, wherein the detection comprises observing more than one signal pulse per nucleotide incorporation event.

76. The method of claim **73**, wherein the mixture is reacted in a zero mode waveguide or nanohole.

77. The method of claim **73**, wherein the modified recombinant polymerase exhibits an increased nucleotide residence time as compared to the parental polymerase.

78. The method of claim **73**, wherein the modified recombinant polymerase exhibits increased processivity as compared to the parental polymerase.

79. A method of making a modified recombinant DNA polymerase, the method comprising:

mutating a polymerase of interest; and

selecting resulting modified polymerases for a property selected from

i) increased branching fraction; and

ii) altered translocation.

80. The method of claim **79**, wherein mutating the polymerase of interest comprises structurally modeling the polymerase to identify a feature that may affect branch fraction or altered translocation.

81. The method of claim **79**, wherein mutating the polymerase of interest comprises making a library of modified recombinant polymerases, and wherein selecting the modified polymerases comprises screening the library to identify at least one member exhibiting the property.

82. The method of claim **79**, wherein the polymerase of interest is a Φ29, B103, GA-1, PZA, Φ15, BS32, M2Y, Nf, G1, Cp-1, PRD1, PZE, SF5, Cp-5, Cp-7, PR4, PR5, PR722, L17, T4 or T7 polymerase.

83. The method of claim **79**, wherein the resulting modified recombinant polymerase that exhibits the property also exhibits an increased nucleotide analog residence time as compared to a corresponding wild-type polymerase.

84. The method of claim **79**, wherein the resulting modified recombinant polymerase that exhibits the property also exhibits increased processivity as compared to a corresponding wild-type polymerase.

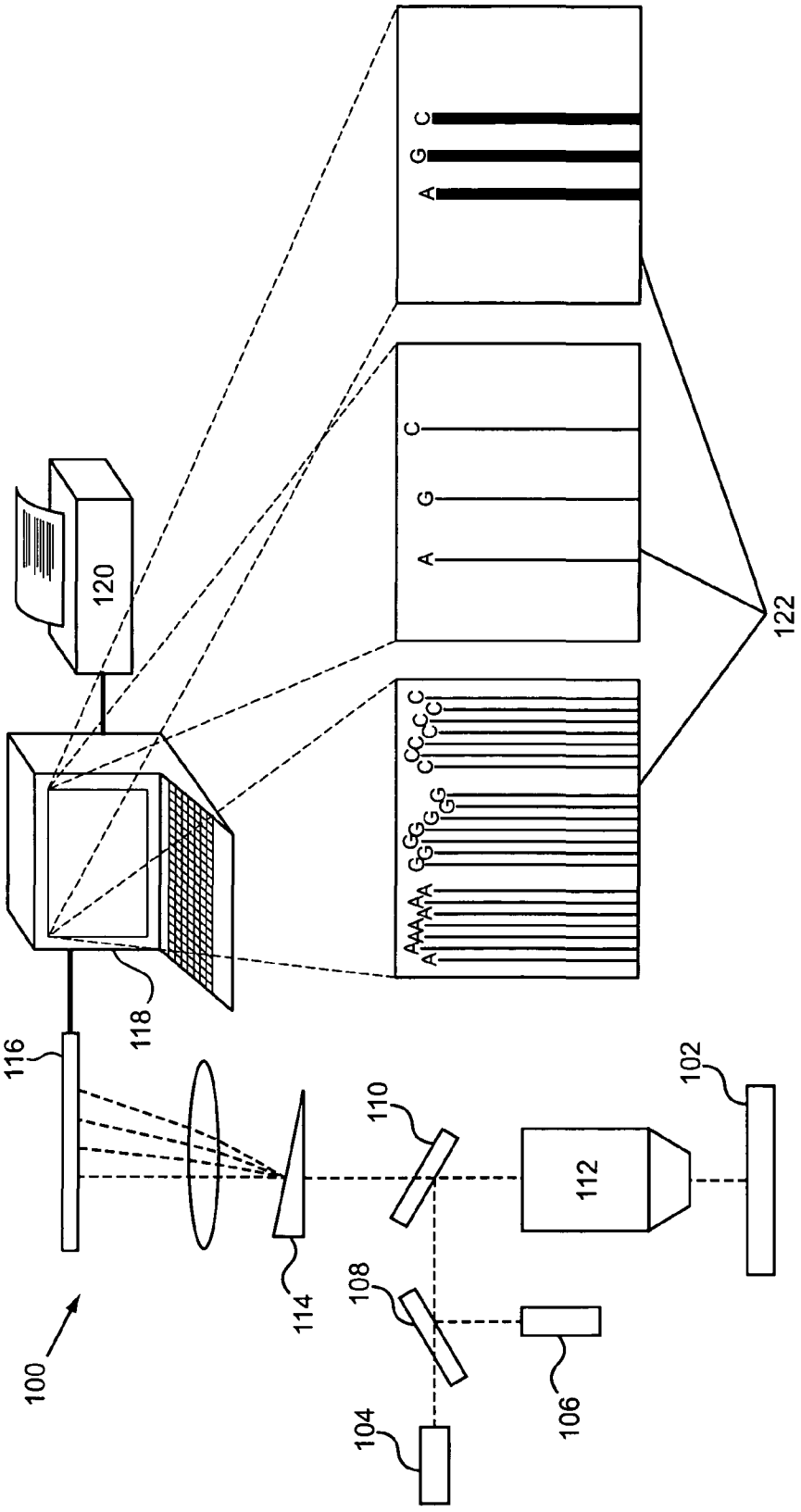

A
G
C
A
G
C
AAAAAA
GGGGGG
CCCCCC
120
122
118
116
114
110
112
102
108
104
106
100

Fig. 1

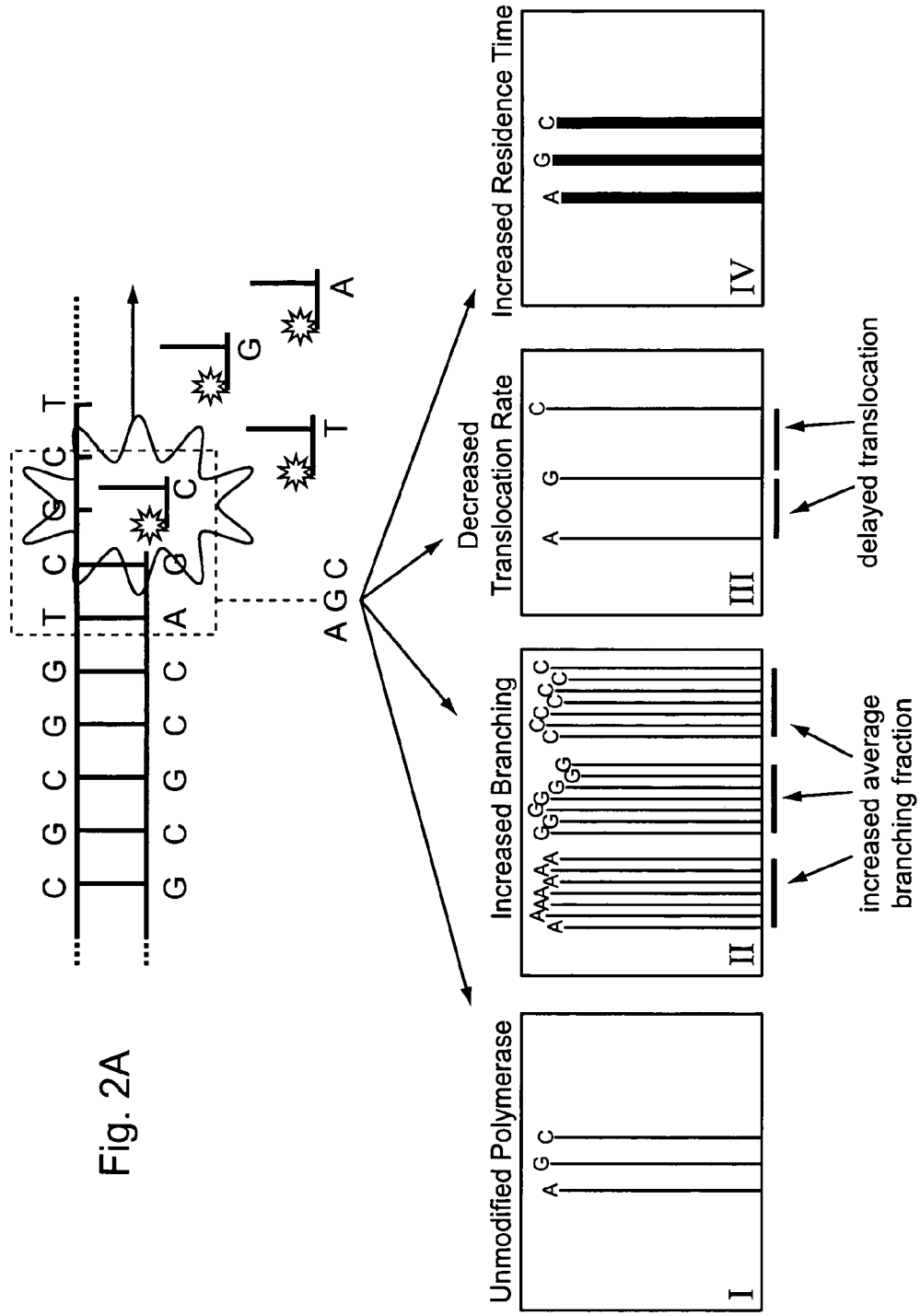

C G C G G T C G C T
G C G C C A G C
A G C
T
G
A
Unmodified Polymerase
I
Increased Branching
II
increased average branching fraction
Decreased Translocation Rate
III
delayed translocation
Increased Residence Time
IV

Fig. 2A

Fig. 2B

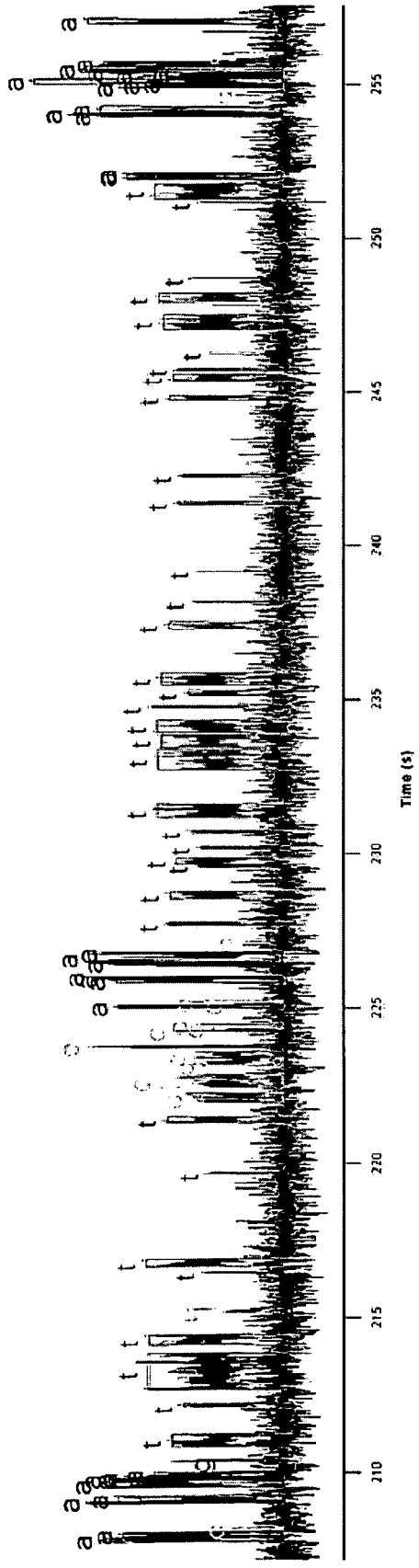


Fig. 3

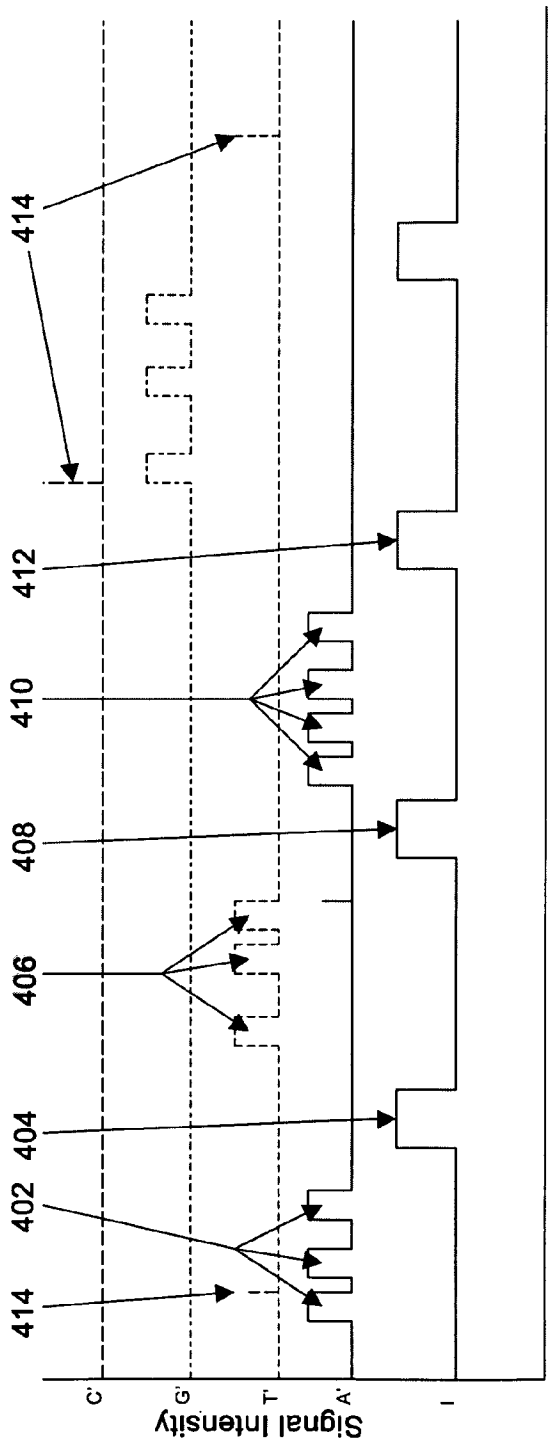


Fig. 4

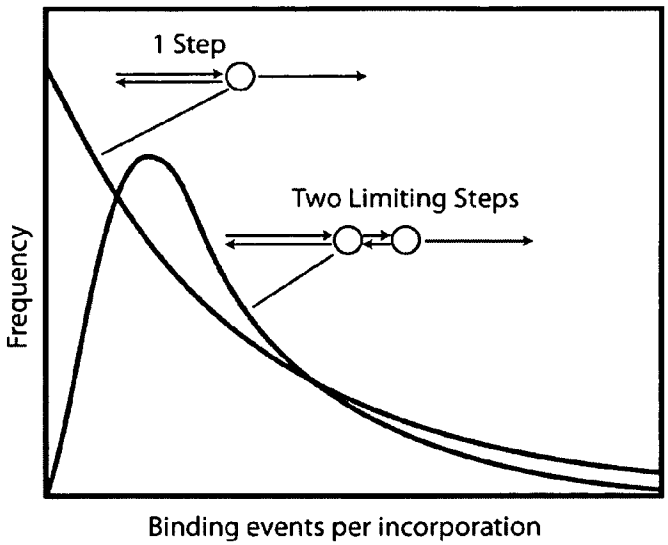

1 Step
Two Limiting Steps
Frequency
Binding events per incorporation

Fig. 5

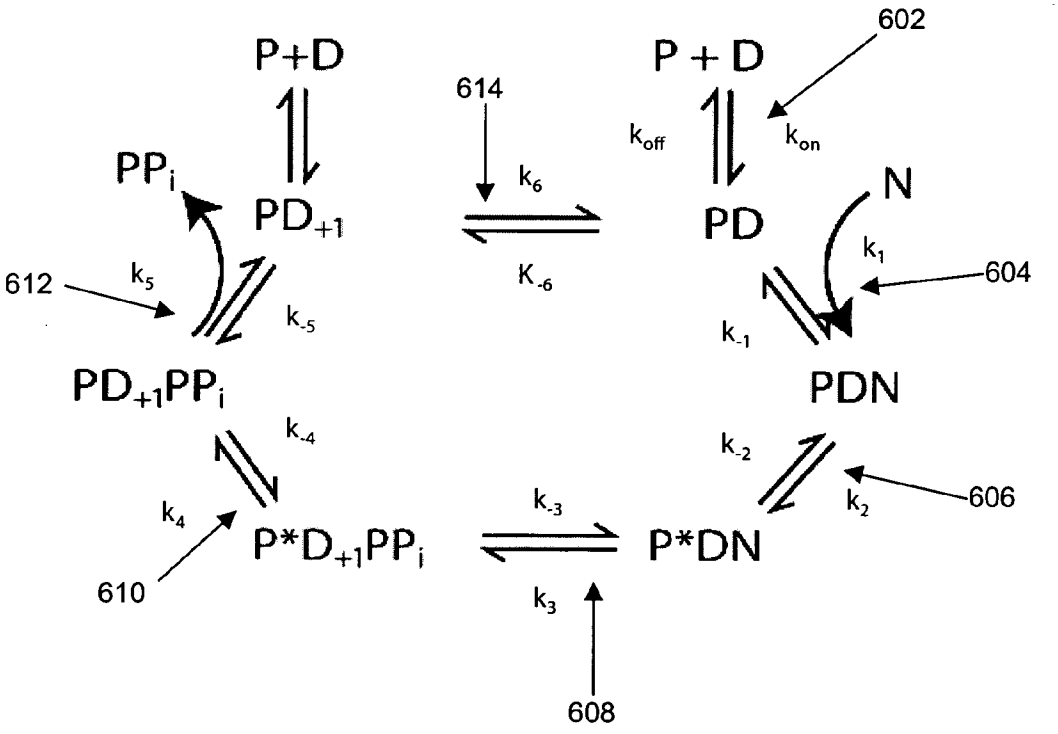

P+D
PP$_i$
PD$_{+1}$
612
k_5
k_{-5}
PD$_{+1}$PP$_i$
k_{-4}
k_4
610
P*D$_{+1}$PP$_i$
614
k_6
K_{-6}
P + D
602
k_{off}
k_{on}
N
PD
k_1
604
k_{-1}
PDN
k_{-2}
k_2
606
k_{-3}
P*DN
k_3
608

Fig. 6

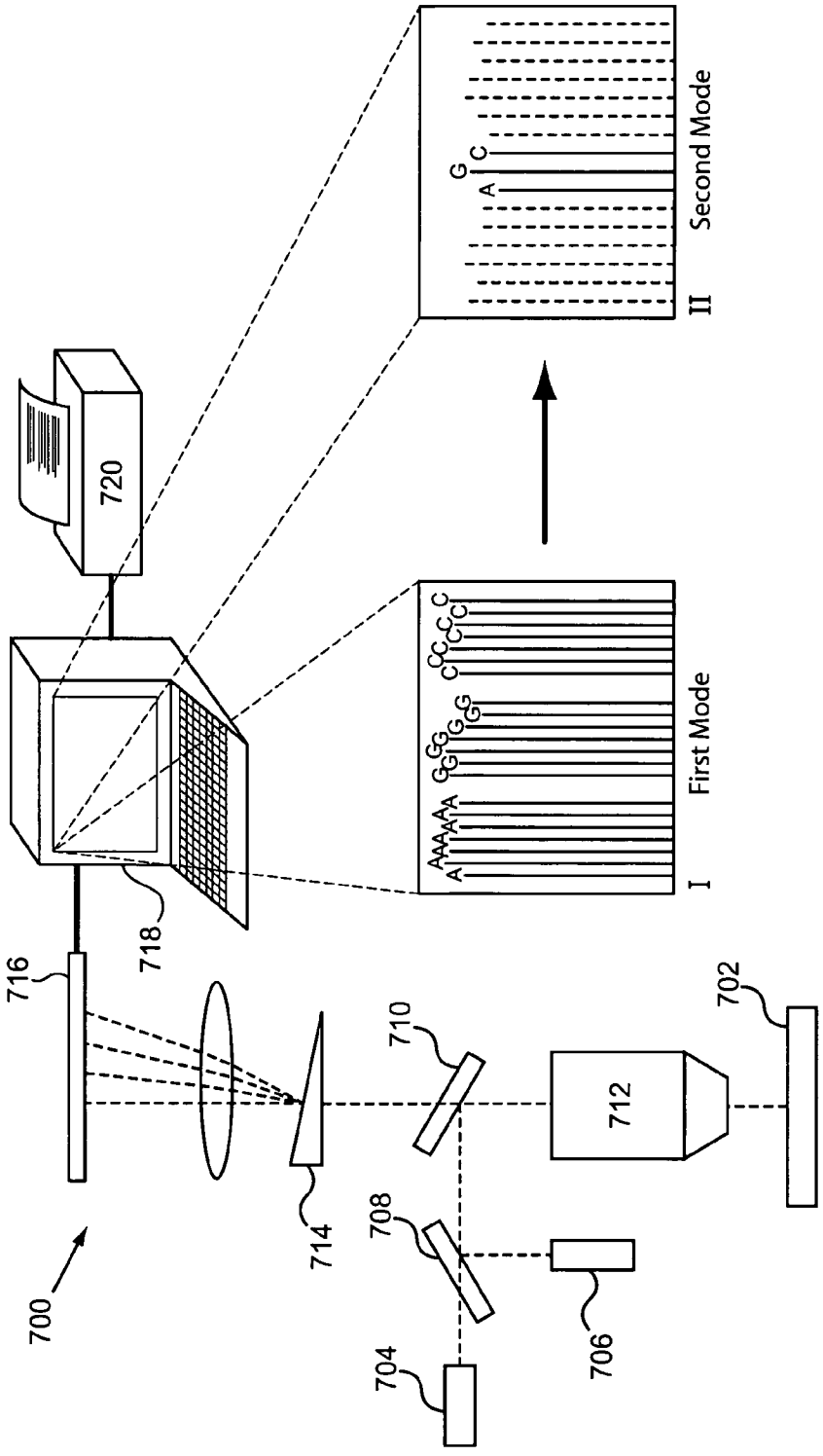

700
702
704
706
708
710
712
714
716
718
720
I
First Mode
II
Second Mode

Fig. 7

Espacenet

Bibliographic data: WO2014153757 (A1) — 2014-10-02

METHOD, SYSTEM, AND COMPUTER READABLE MEDIUM FOR DETERMINING BASE INFORMATION OF PREDETERMINED AREA IN FETAL GENOME

Inventor(s): YIN XUYANG [CN]; JIANG HUI [CN]; CHEN SHENGPEI [CN]; GONG CHUN [CN]; CHEN FANG [CN]; ZHANG CHUNLEI [CN]; PAN XIAOYU [CN] ± (YIN, XUYANG, ; JIANG, HUI, ; CHEN, SHENGPEI, ; GONG, CHUN, ; CHEN, FANG, ; ZHANG, CHUNLEI, ; PAN, XIAOYU)

Applicant(s): BGI SHENZHEN [CN] ± (BGI SHENZHEN)

Classification: **- international:** ***C12Q1/68***
- cooperative: **C12Q1/6876 (EP); G16B20/00 (EP); G16B30/00 (EP);** C12Q2600/156 (EP)

Application number: WO2013CN73375 20130328 Global Dossier

Priority number(s): WO2013CN73375 20130328

Also published as: CN105051208 (A) CN105051208 (B) HK1213945 (A1)

Abstract of WO2014153757 (A1)

A method, a system, and a computer readable medium for determining base information of a predetermined area in a fetal genome are provided. The method for determining base information of a predetermined area in a fetal genome comprises the following steps: obtaining a sequencing result of a DNA sample of a genome of an embryo cell, and a sequencing result of a genome sample of an individual related to embryo heredity; constructing a heredity sketch of the embryo based on the sequencing result of the DNA sample of the genome of the embryo cell, to determine an initial genotype of the embryo; determining haplotypes of parents of the embryo based on the sequencing result of the genome sample of the individual related to the embryo heredity; and using the initial genotype of the embryo as an observation sequence according to the hidden markov model, and determining the base information of the predetermined area in the fetal genome based on the haplotypes of the parents of the embryo.

Patents WO 2014153757

Method, system, and computer readable medium for determining base information of predetermined area in fetal genome

Abstract

A method, a system, and a computer readable medium for determining base information of a predetermined area in a fetal genome are provided. The method for determining base information of a predetermined area in a fetal genome comprises the following steps: obtaining a sequencing result of a DNA sample of a genome of an embryo cell, and a sequencing result of a genome sample of an individual related to embryo heredity; constructing a heredity sketch of the embryo based on the sequencing result of the DNA sample of the genome of the embryo cell, to determine an initial genotype of the embryo; determining haplotypes of parents of the embryo based on the sequencing result of the genome sample of the individual related to the embryo heredity; and using the initial genotype of the embryo as an observation sequence according to the hidden markov model, and determining the base information of the predetermined area in the fetal genome based on the haplotypes of the parents of the embryo.

Classifications

- C12Q1/6876 Nucleic acid products used in the analysis of nucleic acids, e.g. primers or probes

View 3 more classifications

WO2014153757A1
WIPO (PCT)

Download PDF Find Prior Art Similar

Other languages: French, Chinese

Inventor: [redacted]

Worldwide applications

2013 WO CN 2016 HK

Application PCT/CN2013/073375 events

2013-03-28 Application filed by [redacted]

2013-03-28 Priority to PCT/CN2013/073375

2014-10-02 Publication of WO2014153757A1

Info: Patent citations (2), Non-patent citations (1), Legal events, Similar documents, Priority and Related Applications

External links: Espacenet, Global Dossier, PatentScope, Discuss

Claims

Hide Dependent

Claim

A method for determining base information of a predetermined region in an embryo genome, comprising the steps of: obtaining a sequencing result of an embryonic cell genomic DNA sample, and a sequencing result of an embryo genetic related individual genome sample;

Based on the sequencing results of the embryonic cell genomic DNA sample, a genetic sketch of the embryo is constructed to determine the initial genotype of the embryo;

Determining the haplotype of the embryo parent based on the sequencing result of the embryo genetically related individual genomic sample; and based on the hidden Markov model, using the embryo initial genotype as the observation sequence, based on the haplotype of the embryo parent Base information for a predetermined region in the embryonic genome is determined.

2. The method according to claim 1, wherein the embryonic genomic DNA sample is obtained from a microembryonic fetal cell, wherein the microscopic embryonic cell is derived from a blastomere, a blastomere or a fertilized egg.

3. The method according to claim 2, wherein the embryonic genomic DNA sample is obtained from an embryonic single cell.

4. The method according to claim 1, wherein the sequencing result is using a device selected from the group consisting of Illumina-Solexa/Hiseq, ABI-SOLID, Life Technologies-Ion Torrent/Proton, Roche-454, and a single molecule sequencing device. At least one obtained.

5. The method according to claim 1, wherein the genetic sketch of the embryo is constructed by aligning the result of the sequencing of the embryonic cell genomic DNA sample with a reference sequence.

6. The method according to claim 1, wherein the genotype of the embryo genetically related individual is determined by aligning the sequencing result of the embryo genetically related individual genomic sample with a reference sequence;

A haplotype of the parent of the embryo is determined based on the genotype of the embryo-related individual.

7. Method according to claim 5 or 6, characterized in that the reference sequence is a human reference genome.

8. The method of claim 1 wherein the embryo genetically related individual comprises a parent of the embryo and a proband.

9. The method of claim 8 wherein said proband is a child of said embryo parent.

10. The method according to claim 9, wherein determining the base information of the predetermined region in the embryo genome according to the hidden Markov model further comprises:

The initial state probability distribution vector, the probability matrix of the hidden state transition, and the observed sequence probability matrix are constructed; the final state is determined by the Whitby algorithm and the optimal path is traced back to determine the base information of the predetermined region in the embryo genome.

11. The method according to claim 10, wherein the hidden Markov model uses the following parameters: an initial state probability distribution is == 0.5), I-NriNp I = j

隐藏状态转移的概率矩阵为 $A=\{a_{ij}\},(i,j$

eS), where Nr 1 N p

Nr and Np respectively represent the expected number of recombination and single nucleotide polymorphisms, Nr is a natural number, the range is 20-40, and the observed sequence probability matrix is s ≈ {)), G' eS, eV) , where

#sites(L > 0,Must-hom.)

#w'to(L> 0,Must-het. or Must-hom.) b ; (i) = 1 − b ; (0)

#sites (L>0, Must-hom.) is the number of loci that must be homozygous for the offspring. #sites (L>0, Must-hom, or Must-het.) must be homozygous for the offspring. The number of loci and the sum of the number of progeny that must be heterozygous;

P # L ^ (J) = max - 1 ( ) · aij] -bj(K) / F fi NA

The probability of the bureau Shao is Ieil-Wj, and the reverse pointer is

$$\Psi_i(j)=\arg\max \delta_{t-1}(i)\cdot a_{ij}$$

Te{l...N} ^

q N * =argmaxS N (q N)

The final state of recursion is, backtracking the optimal path, determining the base information of the most likely embryo predetermined region is ≈ Ψ ^t(^ +1)

(t = 1,2,3,...,N_1)

12. The method according to claim 1, further comprising: the predetermined region being a site where a genetic polymorphism is known to exist.

13. The method according to claim 12, wherein the genetic polymorphism is at least one selected from the group consisting of single nucleotide polymorphism and STR.

14. A system for determining base information of a predetermined region in an embryonic genome, comprising: a library constructing device adapted to target an embryonic genomic DNA sample and a genomic DNA sample of an embryo genetically related individual, Separate libraries are constructed separately;

a sequencing device coupled to the library construction device and adapted to sequence the sequencing library to obtain sequencing results of embryos and embryo genetically related individuals;

An analysis device, the analysis device being coupled to the sequencing device and adapted to:

Based on the sequencing results of the embryonic genomic DNA sample, a genetic sketch of the embryo is constructed to determine the initial genotype of the embryo;

Determining the haplotype of the embryo parent based on the sequencing result of the embryo genetically related individual genomic sample; and

According to the hidden Markov model, the initial genotype of the embryo is used as an observation sequence, and base information of a predetermined region in the embryo genome is determined based on the haplotype of the embryo parent.

15. The system according to claim 14, further comprising a DNA sample separation device adapted to perform biopsy sampling from the embryo to obtain an amount of embryonic cells from the microembryonic cells. An embryonic genomic DNA sample is extracted, wherein the micro-embryonic cell is derived from a blastomere, a blast, or a fertilized egg.

16. The system according to claim 14, wherein the sequencing device is at least selected from the group consisting of Illumina-Solexa/Hiseq, ABI-SOLID, Life Technologies-Ion Torrent/Proton, Roche-454, and single molecule sequencing devices. One.

17. The system of claim 14 further comprising a comparison device coupled to said sequencing device for comparing said sequencing result to a reference sequence to:

A genetic sketch of the embryo is constructed by aligning the sequencing result of the embryonic cell genomic DNA sample with a reference sequence;

Determining the genotype of the embryonic genetically related individual by comparing the sequencing result of the embryonic genetically related individual genomic sample with a reference sequence;

A haplotype of the parent of the embryo is determined based on the genotype of the embryo-related individual.

18. The system according to claim 14, wherein determining the base information of the predetermined region in the embryo genome according to the hidden Markov model further comprises:

The initial state probability distribution vector, the probability matrix of the hidden state transition, and the observed sequence probability matrix are constructed; the final state is determined by the Whitby algorithm and the optimal path is traced back to determine the base information of the predetermined region in the embryo genome.

19. The system according to claim 18, wherein the hidden Markov model uses the following parameters: an initial state probability distribution is ==0.5), i-NrlNp i = j The probability matrix of the hidden state transition is k =

$\{a_{ij}\},(i,j\in S)$, 其中, a_{ij}

Nr and Np respectively represent the expected number of recombination and single nucleotide polymorphism sites, Nr is a natural number, the range is 20-40, and the observed sequence probability matrix is s ≈ {^)), G' eS, eV) ,

sites(L > 0,Must-hom.)

b, (0)

#w'to(L> 0,Must-het. or Must-hom.) , (1) = 1- b;(0)

#sites (L>0, Must-hom.) is the number of loci that must be homozygous for the offspring. #sites (L>0, Must-hom, or Must-het.) must be homozygous for the offspring. The number of loci and the sum of the number of progeny that must be heterozygous;

P # ^ ^ { J } = max - 1 ( ) · a[j] -b[(K) / F fi NA

The probability of the bureau Shao is Ieil-Wj,

Ψ.. (_/·) = arg max ---·α

The reverse pointer is te{i...N} ^

q N ^ =argmaxS N (q N)

The final state of recursion is, backtracking the optimal path, determining the base information of the most likely embryo predetermined region is ^Ψ [i+1 ^^i) (i = 1,2,3...,N_1)

20. A computer readable medium, wherein the computer readable medium stores instructions, the instructions being adapted to be executed by a processor to:

Based on the sequencing results of the embryonic cell genomic DNA sample, a genetic sketch of the embryo is constructed to determine the initial genotype of the embryo;

Determining the haplotype of the embryo parent based on the sequencing result of the embryo genetically related individual genomic sample; and based on the hidden Markov model, using the embryo initial genotype as the observation sequence, based on the haplotype of the embryo parent Base information for a predetermined region in the embryonic genome is determined.

The computer medium according to claim 20, wherein the genetic sketch of the embryo is constructed by aligning the sequencing result of the embryonic cell genomic DNA sample with a reference sequence.

The computer medium according to claim 20, wherein the genotype of the embryo genetically related individual is determined by comparing the sequencing result of the embryo genetically related individual genomic sample with a reference sequence; The genotype of an embryo-related individual is determined, and the haplotype of the parent of the embryo is determined.

The computer medium according to claim 20, wherein, according to the hidden Markov model, determining base information of a predetermined region in the fetal genome further comprises:

The initial state probability distribution vector, the probability matrix of the hidden state transition, and the observed sequence probability matrix are constructed; the final state is determined by the Whitby algorithm and the optimal path is traced back to determine the base information of the predetermined region in the embryo genome.

24. According to claim 2, the hidden Markov model uses the following parameters: The initial state probability distribution is

$$\pi = \{\pi_i\}, (i \in S, \pi_i = 0.5),$$

I-NrlNp i = j The probability matrix of the hidden state transition is A =

$$\{a_{ij}\}, (i, j \in S),$$

among them,

Nr and Np respectively represent the expected number of recombination and single nucleotide polymorphisms, Nr is a natural number, the range is 20-40, and the observed sequence probability matrix is $^{S\ ^{\bullet}}$ ( )), G' aS, eV) , where

#sites(L > 0,Must-hom.)

#w'to(L> 0,Must-het. or Must-hom.) b ; (i) = 1--- b ; (0)

#sites (L>0, Must-hom.) is the number of loci that must be homozygous for the offspring. #sites (L>0, Must-hom, or Must-het.) must be homozygous for the offspring. The number of loci and the sum of the number of progeny that must be heterozygous;

P # L ^ { J } = max - 1 ( ) · a[j] -b[(K) / F fi NA

The probability of the bureau Shao is Ieil-Wj, and the reverse pointer is

$$\Psi_i(j) = \arg\max \delta_{t-1}(i) \cdot a_{ij}$$

Te{i...N} ^

q N * =argmaxS N (q N)

The final state of recursion is, Backtracking the optimal path, determining the base information of the most likely embryo predetermined region is $^{q\ ^{\bullet}\ \Psi}$ x(^ i + 1) (i = l, 2, 3, ..., N-1)

The computer readable medium according to claim 20, wherein the predetermined area is a site where a genetic polymorphism is known to exist.

The computer readable medium according to claim 25, wherein the genetic polymorphism is at least one selected from the group consisting of a single nucleotide polymorphism and an STR.

Description

Method, system and computer readable medium for determining base information of a predetermined region in an embryonic genome

No technical field

The present invention relates to methods, systems and computer readable media for determining base region information in a predetermined region of an embryonic genome. Background technique

Hereditary diseases are diseases caused by changes in genetic material, which are characterized by congenital, familial, lifelong and hereditary. Hereditary diseases can be divided into three broad categories: monogenic genetic diseases, polygenic genetic diseases, and chromosomal abnormalities. Among them, monogenic diseases are caused by abnormal gene function caused by dominant or recessive inheritance of single disease-causing genes; while polygenic genetic diseases are diseases caused by multiple genetic changes, which will be affected to some extent by external environmental factors. The effects of chromosomal abnormalities include abnormal numbers and structural abnormalities. The most common is Down's syndrome caused by the trisomy of chromosome 21, and the children are characterized by congenital features such as congenital and limb shape abnormalities. Since there is currently no effective treatment for hereditary diseases, only supportive treatment or drug remission can be targeted, which is expensive and imposes a heavy economic and mental burden on society and families. Therefore, it is necessary to test the child's disease state before the child is born, so it is necessary to prevent the work to achieve the purpose of prenatal and postnatal care. For example, in the in vitro reproductive process, it is of great significance to detect the diseased state of the embryo before the embryo is implanted into the uterus. Genetic analysis of gametes or embryos prior to implantation into the uterine cavity involves the implantation of normal embryos that exclude pathogenic factors into the uterus to prevent pregnancy in children with genetic diseases. The detection of the pre-implantation condition of couples with high-risk pregnancy-deficient embryos can undoubtedly reduce the birth of sick children and avoid the many hazards associated with selective termination of pregnancy for pregnant women and families.

However, the current relevant detection methods still need to be improved. Summary of the invention

The present invention has been completed based on the following findings of the inventors:

The inventors have found that the number of cells that can be biopsied for pre-implantation testing is limited. For example, there are only 4-8 cells in the blastomere cultured until the third day, and only 1-2 cells can be taken for biopsy sampling. Even in the embryos that were cultured until the fifth day, only 3-10 cells were taken for the biopsy sampling of the outer trophoblast cells. Due to the limited amount of DNA in a single or a small number of embryonic cells, it is difficult to directly perform a comprehensive genetic variation test. Whole Genome Amplification (WGA) is required to achieve a sufficient amount of DNA to perform a comprehensive analysis of genetic variation. Whole-genome amplification is often biased, resulting in a heterozygous deletion or allele drop out, which poses a risk for the genetic variation associated with genetic disease in embryonic cells.

Therefore, it is necessary to develop a method for locating and correcting errors such as allelic tripping caused by amplification to improve the accuracy of genetic variation detection.

The present invention aims to solve at least one of the technical problems existing in the prior art. Estimating embryos by the method of the invention The ploidy type allows localization of sites where alleles are tripped and can be corrected. The embryo haplotype analysis method can be based on genetic information of family members such as parents and probands, or based on genetic information from other embryonic cells of the same parent.

In one aspect of the invention, the invention proposes a method of determining base information for a predetermined region in an embryonic genome. The method combines lineage inference and statistical algorithms to determine the haplotype of the embryo, and through the genotype of the relevant individual in the family, that is, the transfer of the chromosomal segment, and the statistical algorithm to infer the haplotype. According to an embodiment of the present invention, the method comprises the steps of: obtaining a sequencing result of an embryonic cell genomic DNA sample, and a sequencing result of an embryo genetically related individual genome sample; constructing the inheritance of the embryo based on a sequencing result of the embryonic cell genomic DNA sample Sketching to determine an initial genotype of the embryo; determining a haplotype of the embryo parent based on sequencing results of the embryonic genetically related individual genomic sample; and based on the hidden Markov model, using the initial genotype of the embryo as an observation sequence, based on The haplotype of the embryonic parent determines the base information of a predetermined region in the embryonic genome.

It should be noted that the genomic formation process of the offspring is equivalent to a random recombination of the parental genome (ie, linkage interchange haplotype recombination, and random combination of gametes). Using the haplotype of the embryo as the hidden state, the sequencing data of the whole genome of the embryo single cell is taken as the observations, and the state transition probability is derived by using the Bayesian algorithm with the prior data. Probabilities) , construct observation sequence probabilities and initial state distribution, and then use the Viterbi algorithm to infer the most likely combination of embryonic haplotypes. Thus, according to an embodiment of the present invention, by using a hidden Markov model, for example, by using a Viterbi algorithm, referring to genetic information of an embryo genetically related individual, a nucleic acid sequence of a specific region in the embryo genome can be determined, Therefore, the pre-implantation detection of the genetic information of the embryo genome can be effectively performed. Sites or DNA sequences with low coverage, heterozygous deletion or Allele Drop Out in embryonic cell whole genome amplified DNA sequencing results can be accurately inferred by this method. Therefore, by this method, the position or sequence of the heterozygous deletion or Allele Drop Out can be corrected to make the detection result more accurate and reliable.

In yet another aspect of the invention, the invention provides a system for determining base information for a predetermined region in an embryonic genome. According to an embodiment of the present invention, the system comprises: a library construction device adapted to separately construct a sequencing library for an embryonic genomic DNA sample and a genomic DNA sample of an embryo genetically related individual; a sequencing device, the sequencing The device is coupled to the library construction device and is adapted to sequence the sequencing library to obtain sequencing results of embryos and embryo genetically related individuals; and an analysis device coupled to the sequencing device and adapted to : constructing a genetic sketch of the embryo based on sequencing results of the embryonic genomic DNA sample to determine an initial genotype of the embryo; determining a haplotype of the embryo parent based on sequencing results of the embryonic genetically related individual genomic sample; and The Kraft model uses the embryonic initial genotype as an observation sequence to determine base information of a predetermined region in the embryo genome based on the haplotype of the embryo parent. According to an embodiment of the present invention, an embryo biopsy device may be further included, the embryo biopsy device being adapted to perform biopsy on an in vitro cultured embryo to sample embryonic cells. With the system, the method for determining the predetermined region base information in the embryo genome can be effectively implemented, and the hidden Markov model can be used, for example, by using the Viterbi algorithm to refer to the embryo genetic related individual. Genetic information that identifies the nucleic acid sequence of a particular region of the embryo's genome Therefore, the pre-implantation detection of the genetic information of the embryo genome can be effectively performed, thereby effectively determining the genetic information of the embryo genome.

In another aspect of the invention, the invention also provides a computer readable medium. According to an embodiment of the present invention, a genetic sketch of an embryo is constructed based on a sequencing result of an embryonic cell genomic DNA sample to determine an initial genotype of the embryo; and a haplotype of the embryo parent is determined based on a sequencing result of the embryonic genetic related individual genome sample And determining, based on the hidden Markov model, the initial genotype of the embryo as an observation sequence, based on the haplotype of the parent of the embryo, determining base information of a predetermined region in the embryo genome. With the computer readable medium of the present invention, the stored instructions can be efficiently executed by the processor so as to be able to utilize the hidden Markov model, for example, by using the Viterbi algorithm, based on the sequencing results of the embryonic cells, The genetic information of the embryo-related individuals can determine the nucleic acid sequence of a specific region in the embryo genome, thereby effectively performing pre-implantation detection on the genetic information of the embryo genome.

The additional aspects and advantages of the invention will be set forth in part in the description which follows. DRAWINGS

The above and/or additional aspects and advantages of the present invention will become apparent and readily understood from

1 is a flow chart showing analysis using a hidden Markov model according to an embodiment of the present invention; and FIG. 2 is a schematic structural diagram of a system for determining a predetermined region nucleic acid sequence in an embryo genome according to an embodiment of the present invention. Detailed description of the invention

The embodiments of the present invention are described in detail below, and the examples of the embodiments are illustrated in the drawings, wherein the same or similar reference numerals are used to refer to the same or similar elements or elements having the same or similar functions. The embodiments described below with reference to the drawings are intended to be illustrative only and not to limit the invention.

It should be noted that the terms "first" and "second" are used for descriptive purposes only, and are not to be construed as indicating or implying relative importance or implicitly indicating the number of technical features indicated. Thus, features defining "first", "second" may explicitly or implicitly include one or more of the features. Further, in the description of the present invention, "multiple" means two or more unless otherwise stated.

Method for determining predetermined region information in an embryonic genome

In a first aspect of the invention, the invention proposes a method of determining base information for a predetermined region in an embryonic genome. According to an embodiment of the invention, the method comprises the following steps:

First, the sequencing results of the embryonic cell genomic DNA samples and the sequencing results of the embryo genetically related individual genome samples are obtained. Wherein, the term "embryonic genetically related individual" as used herein refers to an individual that is genetically related to an embryo, for example, an embryonic genetic related individual that can be used according to an embodiment of the present invention. "For the parents of the embryo such as parents, grandparents, grandparents, and other children of the embryonic parents. Said here The other children of the embryo's parents should be understood in a broad sense. They can be either children who have been born, children who are not yet born (fertilized eggs or embryos), embryos that are dead, embryos that are cultured in vitro, or fertilized eggs. As long as the same parents are shared with the embryo to be tested.

According to an embodiment of the present invention, the source of the embryonic cell genomic DNA sample is not particularly limited. According to an embodiment of the present invention, specifically, biopsy sampling may be performed on the embryo, embryonic cells are obtained, and whole cell genome amplification (WGA) is performed on the embryonic cell sample to obtain embryonic cell genomic DNA. The term "embryonic biopsy" as used herein refers to the technique of separating a portion of an embryonic cell from an embryo using a technique or separating a portion of a cell from a fertilized egg/gamete. Wherein, according to an embodiment of the present invention, the embryonic cells may be derived from any one of a blastomere, a trophoblast, a fertilized egg, and a gamete, and may be a single cell or 2 to 10 cells. In addition, whole-genome amplification of any sample of pregnant women containing embryonic nucleic acids can be performed to obtain embryonic genomic DNA samples. Thus, the embryo's genome can be effectively monitored without affecting the normal development of the embryo. Among them, "Genome-wide Amplification (WGA)" mainly includes Multiple Displacement Amplification (MDA) and PCR-based WGA technology, which can use the self-developed WGA amplification process or commercialization. Kits such as Qiagen's REPLI-g series of kits, Sigma-Aldrich's GenomePlex WGA kit (WGA4), Rubicon Genomics' PicoPlex WGA kit, and GE Healthcare's Illustra Genomiphi WGA kit. After amplifying the DNA sample, a sequencing library is constructed separately for the amplified DNA sample of the embryonic cell and the genomic DNA sample of the embryonic genetic related individual.

For methods and procedures for constructing sequencing libraries for nucleic acid samples, those skilled in the art can appropriately select according to different sequencing technologies. For details of the processes, refer to the procedures provided by manufacturers of sequencing instruments such as Illumina, for example, see Illumina Corporation. Multiplexing Sample Preparation Guide (Part #1 005361; Feb 2010) or Paired-End Sample Prep Guide (Part #1 005063; Feb 2010), which is incorporated herein by reference. According to an embodiment of the present invention, a method and apparatus for extracting a nucleic acid sample from a biological sample are also not particularly limited, and can be carried out using a commercially available nucleic acid extraction kit.

After constructing the sequencing library, the sequencing library is applied to the sequencing instrument, the sequencing library is sequenced, and the corresponding sequencing result is obtained, and the sequencing result is composed of a plurality of sequencing data. The method and apparatus that can be used for sequencing according to embodiments of the present invention are not particularly limited, and include, but are not limited to, dideoxy chain termination method; preferably high-throughput sequencing methods, whereby high utilization of these sequencing devices can be utilized The characteristics of flux and deep sequencing further improve sequencing efficiency. Thereby, the subsequent analysis of the sequencing data, especially the accuracy and accuracy of the statistical test analysis, can be improved. The high throughput sequencing methods include, but are not limited to, second generation sequencing techniques or single molecule sequencing techniques. The second generation sequencing platform (Metzker MI.. Sequencing technologies-the next generation. Nat Rev Genet. 2010 Jan; l(l): 31-46) includes, but is not limited to, Illumina-Solexa (GATM, HiSeq2000TM, Miseq, etc.). ;), ABI-SOLID, Life Technologies-Ion Torrent/Proton and Roche-454 (lobe acid sequencing) sequencing platforms; single molecule sequencing platforms (technologies) including but not limited to Helicos's true single molecule sequencing technology (True Single Molecule) DNA sequencing) , Pacific Biosciences single molecule real-time (SMRTTM), and nanopore sequencing technology from Oxford Nanopore Technologies (Rusk, Nicole (2009-04-01). Cheap Third-Generation Sequencing. Nature Methods 6 (4): 244-245). With sequencing With the continuous evolution of technology, those skilled in the art will appreciate that other sequencing methods and devices can be used for whole genome sequencing. According to a specific example of the present invention, the whole genome sequencing library can be sequenced using at least one selected from the group consisting of Illumina-Solexa, ABI-SOLID, Life Technologies-Ion Torrent/Proton, Roche-454, and a single molecule sequencing device. .

According to an embodiment of the present invention, after the sequencing result is obtained, a genetic sketch of the embryo can be constructed based on the result of the sequencing of the embryonic cell genomic DNA sample to determine the initial genotype of the embryo. Based on the sequencing results of the embryonic genetically related individual genome samples, the haplotypes of the embryo parents are determined. According to an embodiment of the present invention, a genetic sketch of the embryo is constructed by aligning the sequencing result of the embryonic cell genomic DNA sample with a reference sequence. Determining the genotype of the embryo genetically related individual by comparing the sequencing result of the embryo genetically related individual genomic sample with a reference sequence; and determining a single of the embryo parent based on the genotype of the embryo genetically related individual Ploid type. According to an embodiment of the invention, a known human reference genome can be used as a reference sequence. For example, according to an embodiment of the invention, the human reference genome used is NCBI 36.3, HG18. Further, according to an embodiment of the present invention, the method of performing the comparison is not particularly limited. According to a particular embodiment of the invention, SOAP can be used for comparison.

Finally, using the initial genotype of the embryo as an observation sequence, based on the haplotype of the embryo parent, the base information of the predetermined region in the embryo genome is determined according to the hidden Markov model.

The term "predetermined region" as used herein shall be taken broadly to mean any region of a nucleic acid molecule that contains a site at which a predetermined event may occur. For SNP analysis, it can refer to the region containing the SNP site. For analysis of chromosome aneuploidy, the predetermined region refers to the full length or portion of the chromosome to be analyzed, i.e., all sequencing data from that chromosome is selected. The method of selecting the sequencing data from the corresponding region from the sequencing results can be not particularly limited. According to an embodiment of the present invention, sequencing data from a

predetermined region can be obtained by aligning all of the obtained sequencing data with a known nucleic acid reference sequence. Further, according to an embodiment of the present invention, the predetermined area may also be a plurality of discrete points on the basis group that are discontinuous. According to an embodiment of the present invention, the type of reference sequence that can be used is not particularly limited and may be any known sequence containing a region of interest.

According to an embodiment of the present invention, base information of the predetermined region can be determined based on a hidden Markov model based on a result of sequencing of the embryonic cells in combination with genetic information of an embryo genetically related individual. According to an embodiment of the present invention, base information of a specific region in the embryo genome can be determined by means of a hidden Markov model using a Viterbi algorithm. Thereby, the pre-implantation detection of the genetic information of the embryo genome can be effectively performed.

Referring now to Figure 1, the principle of analysis using the Whitby algorithm by means of a hidden Markov model is described in detail: As mentioned above, the term "embryonic genetically related individual" as used herein refers to genetically An individual having a relationship with an embryo, for example, an embryonic genetic related individual that can be used according to an embodiment of the present invention is a parent of an embryo such as a parent, a grandparent, a grandparent, and other children of an embryonic parent. The other children of the embryo parents mentioned here should be understood in a broad sense. They can be either children born or unborn children (embryo or fertilized eggs), dead embryos, or embryos cultured in vitro. Or fertilized eggs, as long as they share the same parents as the embryos to be tested.

Thus, the genomic formation process of the offspring is equivalent to a random recombination of the parental genome (ie, linkage-interchange haplotype recombination, and random combination of gametes). Using the embryo's haplotype as a hidden state, the embryo Sequencing data after single-cell whole genome amplification is used as observations, and the transition probabilities are derived by using prior data, and the observation symbol probabilities and initial state probability distribution (initial state probability distribution) are constructed. Distribution), and then the most probable combination of embryonic haplotypes can be inferred by the Viterbi algorithm. Thus, according to an embodiment of the present invention, by using a hidden Markov model, for example, by using a Viterbi algorithm, referring to genetic information of an embryo genetically related individual, a nucleic acid sequence of a specific region in the embryo genome can be determined, Therefore, the pre-implantation detection of the genetic information of the embryo genome can be effectively performed.

The detailed analysis steps are as follows:

Mark:

I. The number of sites that need to be detected is N.

II. Observation sequence;

III. The hidden state set is: "— 1", , which defines which parent's chromosome is passed on to the offspring. 0 represents the chromosome that was inherited to the first i-positive, and 1 represents the chromosome that was not inherited to the proband.

IV. The observation state set is: '.. 1", ^ , 1 means that the chromosome of the inheritance to the proband is consistent with the genotype of the embryo genetic sketch, and 0 represents inconsistency. The first step is to construct the initial state probability distribution vector, haploid recombination. Transfer matrix and observation sequence probability matrix

I. The initial state probability distribution is recorded as ^^^) ^, ¼ 0 ' 5),

In the absence of reference data, it can be assumed that each hidden state is equally likely to occur.

II. The haploid recombination transfer matrix is recorded as A ~ ( ), (² , ⁶ , the probability of hidden state transition. I-Nr I Np I = j

Nr and Np respectively represent the expected number of recombination and single nucleotide polymorphism sites. According to the prior data Nr, a natural number between 20 and 40 can be taken to calculate the transition probability of the hidden state, that is, calculate the base combination of each base. The probability that a haplotype will recombine.

III. Record the probability matrix of the observed sequence as β = (^, G≡S, k eV) 5 The probability distribution here must be homozygous (Must-hom) or must be heterozygous (Must-het). To estimate.

#sites(L > 0, Must-hom.)

¥0)

#w'to(L> 0, Must-het. or Must-hom.) bl.(l) = 1— (0) The second step is to construct a local probability matrix and a reverse pointer. I...N) i (j) = argmaxS t – 1 (i)-a

Definition Reverse pointer: {i-N} The third step, determine the final state state, and backtrack the optimal path

q N * =argmaxS N (q N)

Determine the final state, . Backtracking the optimal path, that is, the most probable embryo single genotype qt fourth step, outputting the result, according to an embodiment of the present invention, determining the base information of the predetermined region in the embryo genome according to the hidden Markov model further includes:

The initial state probability distribution vector, the probability matrix of the hidden state transition, and the observed sequence probability matrix are constructed; the final state is determined by the Whitby algorithm and the optimal path is traced back to determine the base information of the predetermined region in the embryo genome.

According to a specific embodiment, the hidden Markov model uses the following parameters: The initial state probability distribution is = {^, £5 ¼ = 0 . 5 }

The probability matrix of the hidden state transition is ^ =

$$\{a_{ij}\},(i,j \in S)$$，其中，$$a_{ij} = \begin{cases} 1 - Nr / Np & i = j \\ Nr / Np & i \neq j \end{cases}$$

eS where,

Nr and Np respectively represent the expected number of recombination and single nucleotide polymorphisms, Nr is a natural number, the range is 20-40, and the observed sequence probability matrix is $^{s\ n}\{\}$, {i' eS, eV} , where

#sites(L > 0,Must-hom.)

sites (L > 0, Must-het. or Must-hom.) b.(i) = i-b.(o)

#sites (L>0, Must-hom.) is the number of loci that must be homozygous for the offspring. #sites (L>0, Must-hom, or Must-het.) must be homozygous for the offspring. The number of loci and the sum of the number of progeny must be the number of heterozygous loci; the probability of the local swell is leil-Wj, and the reverse pointer is

$$\Psi_i(j) = \arg\max \delta_{t-1}(i) \cdot a_{ij}$$

Te{l...N} ^

$q_N^* = \arg\max S_N(q_N)$

The final state is recursively based on the Whitby algorithm.

Backtracking the optimal path, determining the base letter of the most likely embryo predetermined area, the information is $= \Psi\ ^{\wedge}(^{\wedge} ^{t}\ )$ (i = 1, 2, 3, ..., N_1) Here, the terms "local probability ( )" and "reverse pointer Ψ ; {) " are used in the classic definition of the Viterbi algorithm. For a detailed description of the definition of this parameter, see Lawrence R. Rabiner, PROCEEDINGS OF THE IEEE, Vol. 77, No. 2, February 1989, which is incorporated herein in its entirety by reference.

Thereby, the sequence of the embryonic genome can be efficiently analyzed. Compared with other existing pre-implantation detection techniques, this method has the following technical advantages, mainly in terms of accuracy and the amount of genetic information available:

1) According to an embodiment of the present invention, both the paternal and maternal sites of the embryo can be well detected, the detection accuracy can be as high as 95% or more, and multiple variant types can be detected, and the disease is expanded. The scope of the test.

2) According to an embodiment of the present invention, for some sites or DNAs with low coverage, heterozygous deletion or Allele Drop Out in embryonic single cell or micro cell whole genome amplified DNA sequencing results The sequence can be accurately inferred by this method. Therefore, by this method, the position or sequence of the heterozygous deletion or Allele Drop Out can be corrected to make the detection result more accurate and reliable.

3) According to the embodiment of the present invention, genetic disease mapping can be performed. For some chain related diseases, it can be directly inferred from the information of other sites, and the amount of information available at one time is large, which is more instructive for clinical detection.

In addition, according to an embodiment of the present invention, the method for determining base information of a predetermined region in an embryo genome is not limited to a genetic polymorphism site such as SNP or STR, and may be used for all genetic polymorphism sites. Applicable, and multiple sites can be used simultaneously to verify each other. System for determining predetermined regional information in the embryonic genome

In yet another aspect of the invention, the invention provides a system for determining a predetermined region of a nucleic acid sequence in an embryonic genome. According to an embodiment of the present invention, referring to FIG. 2, the system 1000 can include: a library construction device 100, a sequencing device 200, and an analysis device 400.

According to an embodiment of the invention, library construction device 100 is adapted to construct a sequencing library separately for embryonic genomic DNA samples and genomic DNA samples of embryo genetically related individuals. According to an embodiment of the invention, the sequencing device 200 is coupled to the library construction device 100 and is adapted to sequence the sequencing library to obtain sequencing results for embryos and embryo genetically related individuals. According to an embodiment of the present invention, a DNA sample separation device and a DNA amplification device (not shown) may be further included, and the DNA sample separation device is adapted to perform biopsy sampling on the embryo to obtain embryonic cells. The embryonic cells may be derived from any one of a blastomere, a trophoblast, a fertilized egg, and a gamete, and may be a single cell, or may be a cell, such as 2-10 cells, or may contain any embryo. A sample of pregnant women with nucleic acids. The DNA amplification device is adapted for whole genome amplification of embryonic cells obtained by biopsy sampling to obtain a sufficient amount of DNA for subsequent detection analysis. Thus, the embryo's genome can be effectively monitored without affecting the normal development of the embryo. Methods and procedures for whole-genome amplification, including Multiple Displacement Amplification (MDA) and PCR-based WGA technology, can be used in the self-developed WGA amplification process, or can use commercial reagents. Boxes such as Qiagen's REPLI-g series of kits, Sigma-Aldrich's GenomePlex WGA kit (WGA4), Rubicon Genomics' PicoPlex WGA kit, GE-Healthcare's illustra GenomiPhi WGA kit, and more. For methods and procedures for constructing sequencing libraries for nucleic acid samples, those skilled in the art can appropriately select according to different sequencing technologies. For details of the processes, refer to the procedures provided by manufacturers of sequencing instruments such as Illumina, for example, see Illumina Corporation. Multiplexing Sample Preparation Guide (Part #1 005361; Feb 2010) or Paired-End SamplePrep Guide (Part #1 005063; Feb 2010), which is incorporated herein by reference. According to an embodiment of the present invention, the method and apparatus for extracting a nucleic acid sample from a biological sample are also not particularly limited, and can be carried out using a commercially available nucleic acid extraction kit. The method and apparatus that can be used for sequencing according to embodiments of the present invention are not particularly limited, and include, but are not limited to, dideoxy chain termination method; preferably high-throughput sequencing methods, whereby high utilization of these sequencing devices can be utilized The characteristics of flux and deep sequencing further improve sequencing efficiency. Thereby, the subsequent analysis of the sequencing data, especially the accuracy and accuracy of the statistical test analysis, can be improved. The high throughput sequencing methods include, but are not limited to, second generation sequencing techniques or single molecule sequencing techniques. The second generation sequencing platform (Metzker ML. Sequencing technologies-the next generation. Nat Rev Genet. 2010 Jan; 11(1): 31-46) includes but is not limited to Illumina-Solexa (GATM, HiSeqTM, Miseq, etc.)), ABI-SOLID, Life Technologies-Ion Torrent/Proton, and Roche-454 (pyrophosphate sequencing) sequencing platforms; single-molecule sequencing platforms (technologies) including but not limited to Helicos' true single-molecule sequencing technology (True Single Molecule DNA sequencing)), Pacific Biosciences' single molecule real-time (SMRTTM), and nanopore sequencing technology from Oxford Nanopore Technologies (Rusk, Nicole (2009-04-01). Cheap Third-Generation Sequencing. Nature Methods 6 (4): 244-245). As sequencing technologies continue to evolve, those skilled in the art will appreciate that other sequencing methods and devices can be used for whole genome sequencing. According to a specific example of the invention, the whole genome sequencing library can be sequenced using at least one selected from the group consisting of Illumina-Solexa, ABI-SOLID, Life Technologies-Ion Torrent/Proton, Roche-454, and single molecule sequencing devices.

Alignment device 300 may also be included in accordance with an embodiment of the present invention. According to an embodiment of the invention, the comparison device 300 is coupled to the sequencing device 200 and is adapted to align the resulting sequencing results with a reference sequence for comparison of the sequencing results of the embryonic cell genomic DNA samples with a reference sequence Constructing a genetic sketch of the embryo; determining a genotype of the embryo genetically related individual by comparing the sequencing result of the embryo genetically related individual genomic sample with a reference sequence; and a gene based on the embryo genetically related individual Type, determining the haplotype of the parent of the embryo. According to an embodiment of the present invention, the type of reference sequence that can be used is not particularly limited and may be any known sequence containing a region of interest. According to an embodiment of the present invention, a known human reference genome can be used as a reference sequence. For example, according to an embodiment of the invention, the human

reference genome used is NCBI 36.3, HG18. Further, according to an embodiment of the present invention, the method of performing the comparison is not particularly limited. According to a particular embodiment of the invention, SOAP can be used for comparison.

According to an embodiment of the present invention, the analyzing device 400 is adapted to: construct a genetic sketch of the embryo based on the sequencing result of the embryonic cell genomic DNA sample, to determine an initial genotype of the embryo; and sequence the result based on the genome sample of the embryo genetically related individual Determining a haplotype of the embryonic parent; and determining, based on the hidden Markov model, the base information of the predetermined region in the embryo genome based on the initial genotype of the embryo as an observation sequence, based on the haplotype of the parent of the embryo.

According to an embodiment of the present invention, base information of a predetermined region in an embryo genome is determined according to a hidden Markov model Further includes:

The initial state probability distribution vector, the probability matrix of the hidden state transition, and the observed sequence probability matrix are constructed; the final state is determined by the Whitby algorithm and the optimal path is traced back to determine the base information of the predetermined region in the embryo genome.

According to a specific embodiment, the hidden Markov model uses the following parameters:

The initial state probability distribution is = $^{E5\ i_A\ =0\ \_\ 5}$),

隐藏状态转移的概率矩阵为 $A=\{a_{ij}\},(i,j\in S)$，其中，$a_{ij}=\begin{cases}1-Nr/Np & i=j\\ Nr/Np & i\neq j\end{cases}$

Nr, Np respectively represent the expected number of recombination and single nucleotide polymorphism sites, Nr can take a natural number between 20-40, and observe the sequence probability matrix as $^{\&\ \alpha}$ {^))G' eS eV) , where

sites(L > 0,Must-hom.)

b, (0)

sites (L > 0 Must-het. or Must-hom.) , (1) = 1— b; (0)

#sites (L>0, Must-hom.) is the number of loci that must be homozygous for the offspring. #sites (L>0, Must-horn, or Must-het.) must be homozygous for the offspring. The number of loci and the sum of the number of progeny that must be heterozygous;

P $_{\#}$ ^ ^ (J) = max - 1 ( ) · ai]} -b[(K) $_{/\ F\ ff\ NA}$

The probability of the bureau Shao is ieil-Wj,

Xi $_{;}$ (j) $^{=\ ar}$ g $^{maXi5\ i}$ $_{;}$ -i($^{-3}/_{\wedge\ ii}$

The reverse pointer is '

q $_{N}$ * =argmaxS $_{N}$ (q $_{N}$)

The final state of recursion is , backtracking the optimal path, determining the base information of the most likely predetermined region of the embryo as

(i = 123... N_1 The terms local probability and inverse pointer 这里 are used here , ( ) are the classic definitions of the Viterbi algorithm. For a detailed description of the definition of this parameter, see Lawrence R. Rabiner, PROCEEDINGS OF THE IEEE, Vol. 77, No. 2, February 1989, the entire contents of which are incorporated herein by reference.

The data analysis section has been described in detail above, and is of course applicable to a system for determining a predetermined region nucleic acid sequence in the embryonic genome. No longer.

Thus, with the system, the method for determining a predetermined region nucleic acid sequence in the embryo genome can be effectively implemented, and the embryonic genome can be determined by, for example, a Viterbi algorithm decoding by means of a hidden Markov model. The base information of a specific region, whereby the genetic information of the embryo genome can be effectively detected before implantation.

Further, according to an embodiment of the present invention, the predetermined region is a site where the genetic polymorphism is known to exist, and the genetic polymorphism is at least one selected from the group consisting of a single nucleotide polymorphism and an STR. The term "connected" as used herein shall be understood broadly and may be either directly connected or indirectly connected as long as the above functional connections are achieved.

It should be noted that those skilled in the art can understand that the features and advantages of the method for determining the predetermined region nucleic acid sequence in the embryo genome described above are also suitable for the system for determining the predetermined region nucleic acid sequence in the embryo genome, which is convenient for description, More details. Computer readable medium

In yet another aspect of the invention, the invention provides a computer readable medium. According to an embodiment of the invention, instructions are stored on the computer readable medium, the instructions being adapted to be executed by a processor to:

Based on the sequencing results of the embryonic cell genomic DNA sample, a genetic sketch of the embryo is constructed to determine the initial genotype of the embryo;

Determining the haplotype of the embryo parent based on the sequencing result of the genetic reference sample of the embryo related to the embryo; and determining the embryo genome by the initial genotype of the embryo as the observation sequence according to the hidden Markov model, based on the haplotype of the parent of the embryo Base information of a predetermined area.

Thus, with the computer readable medium, the method described above can be effectively implemented, so that base information of a specific region in the embryo genome can be determined by, for example, a Viterbi algorithm using a hidden Markov model. Thus, the pre-implantation detection of the genetic information of the embryonic genome can be effectively performed.

According to an embodiment of the present invention, determining the base information of the predetermined region in the embryo genome according to the hidden Markov model further includes:

The initial state probability distribution vector, the probability matrix of the hidden state transition, and the observed sequence probability matrix are constructed; the final state is determined by the Whitby algorithm and the optimal path is traced back to determine the base information of the predetermined region in the embryo genome.

According to a specific embodiment, the hidden Markov model uses the following parameters:

The initial state probability distribution is = £ S ιA = 0 . 5) , I - Nr i Np i = j

A = { aij } , (i, je S) ^ where Nr / Np i≠ j hides the probability matrix of state transitions as

Nr and Np respectively represent the expected number of recombination and single nucleotide polymorphism sites, Nr can take a natural number between 20-40, and observe the sequence probability matrix for its towel.

sites (L > 0, Must-horn.)

b, (o)

sites (L > 0, Must-het. or Must-horn.) ; (1) = 1 — b; (0)

#sites (L>0, Must-horn.) is the number of loci that must be homozygous for the offspring. #sites (L>0, Must-horn, or Must-het.) must be homozygous for the offspring. The number of loci and the sum of the number of progeny that must be heterozygous; The probability of the bureau Shao is lell-Wj

Ψ, (_/·) = arg max a

The reverse pointer is te{l...N} ^

q_N * =argmaxS_N (q_N)

After recursively, the final state is the backtracking optimal path, and the most likely embryo predetermined region is determined. The base information of the domain is ˇ Ψ ^ι(^ + 1) (l = 1, 2, 3, ..., N_1) The terms local probability ( ) and inverse pointer Ψ ; ( ) are used in the classic definition of the Viterbi algorithm. For a detailed description of the definition of this parameter, see Lawrence R. Rabiner, PROCEEDINGS OF THE IEEE, Vol.77, No .2, February 1989, incorporated herein by reference in its entirety.

The data analysis section has been described in detail above, and is of course applicable to a system for determining a predetermined region nucleic acid sequence in the embryonic genome. No longer.

Thus, with the system, the method for determining a predetermined region nucleic acid sequence in the embryo genome can be effectively implemented, and the specificity of the embryo genome can be determined by, for example, a Viterbi algorithm using a hidden Markov model. The base information of the region, thereby enabling pre-implantation detection of the genetic information of the embryonic genome.

Furthermore, according to an embodiment of the present invention, the predetermined region is a site where the genetic polymorphism is known to exist, and the genetic polymorphism is at least one selected from the group consisting of a single nucleotide polymorphism and an STR.

For the purposes of this specification, a "computer-readable medium" can be any apparatus that can contain, store, communicate, propagate, or transport the program for use by the instruction execution system, apparatus, or device, or in conjunction with the instruction execution system, apparatus, or device. More specific examples (non-exhaustive list) of computer readable media include the following: electrical connections (electronic devices) having one or more wires, portable computer disk cartridges (magnetic devices), random access memory (RAM), Read only memory (ROM), erasable editable read only memory (EPROM or flash memory), fiber optic devices, and portable compact disk read only memory (CDROM). In addition, the computer readable medium may even be a paper or other suitable medium on which the program can be printed, as for example by optical scanning of paper or other media, followed by editing, interpretation or other suitable if necessary The method proceeds to obtain the program electronically and then store it in computer memory.

It should be understood that portions of the invention may be implemented in hardware, software, firmware or a combination thereof. In the above-described embodiments, multiple steps or methods may be implemented in software or firmware stored in a memory and executed by a suitable instruction execution system. For example, if implemented in hardware, as in another embodiment, it can be implemented with any one or combination of the following techniques well known in the art: having logic gates for implementing logic functions on data signals Discrete logic circuits, application specific integrated circuits with suitable combinational logic gates, programmable gate arrays (PGAs), field programmable gate arrays (FPGAs), etc.

A person skilled in the art can understand that all or part of the steps carried by the method of the foregoing embodiment can be completed by a program to instruct related hardware, and the program can be stored in a computer readable storage medium. The program, when executed, includes one or a combination of the steps of the method embodiments.

In addition, each functional unit in each embodiment of the present invention may be integrated into one processing module, or each unit may exist physically separately, or two or more units may be integrated into one module. The above integrated modules can be implemented in the form of hardware or in the form of software functional modules. The integrated modules, if implemented in the form of software functional modules and sold or used as separate products, may also be stored in a computer readable storage medium. The solution of the present invention will be explained below in conjunction with the embodiments. Those skilled in the art will appreciate that the following examples are merely illustrative of the invention and are not to be considered as limiting the scope of the invention. In the examples, the specific techniques or conditions are not indicated, according to the techniques or conditions described in the literature in the field (for example, refer to J. Sambrook et al., Huang Peitang et al., Molecular Cloning Experimental Guide, Third Edition, Science Press) or in accordance with the product manual. The reagents or instruments used are not specified by the manufacturer, and are conventional products that are commercially available, for example, from Illumina.

General method:

Referring to FIG. 1, the main steps of the embodiment of the present invention include:

1) Obtain single-cell or micro-cell samples, third-day blastomere cells or fifth-day blastocyst trophoblast biopsy sampling, or fertilized egg sampling, or other single-cell or micro-cell sampling.

2) Whole-genome amplification of single-cell or micro-cell samples, and the amplified product DNA is constructed according to the requirements of the next-generation sequencing platform and sequenced.

3) Single cell or micro cell sequencing data is filtered and compared to the human genome reference sequence.

4) Preliminary detection of genetic variation such as single nucleotide polymorphisms and deletions in embryonic cells based on the alignment results.

5) Construct a genetic sketch of the embryonic genome to initialize the embryonic genotype.

6) Collecting embryonic genetically related individuals, such as family members such as parents and probands, and/or grandparents, grandparents, etc., or other embryo samples from the same parent, extracting genomic DNA, constructing a library according to the requirements of a new generation sequencing platform, and performing Sequencing.

7) Embryonic genetically related individuals DNA sequencing data is filtered and compared to the human genome reference sequence.

8) Analyze and determine the genotypes of genetically related individuals such as probands and parents based on the results of the comparison.

9) Inferring the haplotypes of parents and parents using the genotypes of genetically related individuals such as probands and parents.

10) Using the hidden Markov model decoding method, the embryo haplotype was inferred from the parental hapiotype results.

11) Final determination of genetic variation in the embryo.

Referring to FIG. 1, the hidden Markov model used in the information analysis part of the embodiment of the present invention and the detailed analysis steps are as follows:

Mark:

1. The number of sites to be detected is N.

2. Observation sequence:

3. The hidden state set is: Two ^^;^, which defines which parent's chromosome is passed on to the offspring. 0 means inheritance first The chromosome of the witness.

4. The observed state set is: V = {0, 1}. 1 represents the chromosome that is inherited to the proband and is consistent with the genetic sketch of the embryo.

0 means inconsistency. The first step is to construct the initial state probability distribution vector, the haploid recombination transfer matrix and the observed sequence probability matrix.

i. The initial state probability distribution is recorded as; r = {; r ; }, (i∈S, ^ = 0.5) in the absence of reference data, it is assumed that each hidden state is equally likely to occur.

ii. The haploid recombination transfer matrix is recorded as A = , (ε' eS), which is the probability of hidden state transition.

$$a_{ij} = \begin{cases} 1 - Nr / Np & i = j \\ Nr / Np & i \neq j \end{cases}$$

Nr, Np represent the expected number of recombination and single nucleotide polymorphism sites, respectively.

Constructing an observation sequence probability matrix

Record the sequence probability matrix as 8

$$= \{b_i(k)\}, (i \in S, k \in V),$$

The probability distribution here must be homozygous with the offspring.

(Must-hom) or must be a hetero-het site to estimate.

$$b_i(0) = \frac{\#\, sites(L > 0, \text{Must-hom.})}{\quad}, \quad b_i(1) = 1 - b_i(0)$$

sites L > 0, Must-het. or Must-hom. The second step is to construct a local probability matrix, and a reverse pointer.

Define the local probability: St (j) = max [St aij] -bj(K) te{i...N} Define the inverse pointer: Ψ, (j) = arg max 5 $_{i \ldots s}$ (i) - a.. t {i...N} The third step is to determine the final state state and backtrack the optimal path.

Determine the final state, q_N * = arg max S_N (q_N) Backtracking the optimal path, is the most likely embryo single genotype *=ψ_{i+1} ($^{\wedge}{}_{i+1}$) (= i,2,3... .,N-1)

Fourth step, output results Example 1

First, sample collection and processing:

A couple who had given birth to a daughter with severe beta thalassemia (proband), therefore receiving IVF (in vitro fertilization combined with embryo transfer technology) - PGD treatment, hope to have a health that is consistent with the proband's HLA matching baby. After an IVF-PGD treatment cycle, a single-cell biopsy of the blastomere embryo was taken on the third day, and the blastomere after the biopsy was further cultured until the fifth day of the embryonic stage, and the embryo was vitrified for cryopreservation. Single-cell sample of blastomere obtained by biopsy Whole genome amplification (PED7-1 - GWGA), whole genome amplification, using Qiagen's REPLI-g MDA amplification kit, operating strictly according to the kit instructions, amplification products for next-generation sequencing and other analyses. The peripheral blood of the father (PED4-F), mother (PED3-M) and daughter of the proband (PED-Sister) in this family was also collected and genomic DNA was extracted.

After completing IVF-PGD treatment, select the correct embryo to be implanted into the uterus. After successful pregnancy, pregnant women with amniotic fluid (PED7-2-OAF) can be taken after 14 weeks, or embryonic cord blood can be collected, DNA extracted, and sequenced to verify The result of typing of embryonic cells. Embryo single-cell whole-genome amplification products DNA was disrupted to a 200-bp fragment using a CovarisTM interrupter, and the library was constructed according to the requirements of the Illumia® HiSeq2000TM sequencer.

Agilent® Bioanalyzer 2100 and Q-PCR methods for quality control, qualified libraries continue with Illumina®

The HiSeq2000TM sequencer was sequenced and the sequencing strategy was Pair End 90 index (ie, bidirectional 90bp index sequencing;), sequencing depth 30X, whole genome sequencing. The instrument's parameter settings and operating methods are in accordance with the Illumina® operating manual (available at http://www.illumina.com/support/documentation.ilmn). The peripheral blood DNA of the father, mother, and proband daughter, as well as the amniotic fluid or embryonic cord blood DNA of the pregnant woman, was interrupted to a 200 bp fragment by a CovarisTM interrupter, and the library was built according to the requirements of the Illumina® HiSeq2000TM sequencer. The library was captured by All in One chip. For details, see the NimbleGen SeqCap EZ chip capture library description (available at http://www.nimblegen.com/products/seqcap/ez/index.html). The All in One chip is an autonomously designed target region capture chip with a target region including a full exon region, a 1M SNP site region, and an MHC gene region. The chip-captured library was quality-controlled by the Agilent® Bioanalyzer 2100 and Q-PCR methods. The qualified library was further sequenced by the Illumina® HiSeq2000TM sequencer. The sequencing strategy was Pair End 90 index, and the sequencing depth was 30-50X in the target area. The instrument's parameter settings and operating methods are in accordance with the Illumina® operating manual (available at http://www.illumina.com/support/documentation.ilmn). Second, sequencing data analysis:

1. Proband and parental genome DNA sequencing typing:

a) Filter out sequences of low mass values and joint contamination

b) Use Burrows-Wheeler Aligner (BWA, http /bio-bwa.sourceforge.net/)^)!"High quality ó: ill sequence alignment in the human reference genome (Hgl9, NCBI release GRCh37) and removed with SAMtools The amplified sequence was repeated by PCR.

c) Calibrate the alkali matrix mass using the Genome Analysis Toolkit (GATK, http: 〃 www.broadinstitute.org/gatk/) and perform single-base nucleotide polymorphisms and deletions.

2. Construction of genetic sketches of embryonic genomes.

3. Inferring the parent's haplotype based on the proband's genotype:

The haplotype is inferred from the core family, and can be specifically referred to in BL Browning and SR Browning. Am J Hum Genet 84: 210-223. (2009). 4. Inferring the haplotype of embryonic cells based on parental haplotype:

a) Use BWA to compare embryonic single cell sequencing data to the human reference genome (Hgl9, NCBI release GRCh37)

b) Construct an initial state probability vector, and an embryonic haploid recombination transfer matrix, Nr = 30, as described above. c) Count the sequencing information for each site of the embryonic cells and construct an observation sequence probability matrix (as described above).

d) Construct a local probability matrix, and a reverse pointer (as described above).

e) determine the final state and backtrack the optimal path

f) output

g) Based on the results of amniotic fluid genotyping after embryogenesis, the genotyping accuracy of embryonic cells is as follows:

种类		常染色体			性染色体		
		准确数	位点数	准确率	准确数	位点数	准确率
重构前	纯合	175,194	233,794	74.94	2,345	3,318	70.68
	杂合	81,129	81,576	99.45	1,476	1,480	99.72
重构后	纯合	172,968	173,631	99.62	2,329	2,330	99.96
	草图杂合	80,877	81,042	99.80	1,472	1,478	99.59
	更正杂合	58,194	60,194	95.88	967	990	97.68

Note: "Sketch hybrid" means a site that is heterozygous before and after recombination; "Correct heterozygous" means homozygous before reconstitution and heterozygous after recombination.

5, β-thalassaemia and HLA type judgment

a) The genotype of embryonic cells on rs7480526 was successfully restored by parental haplotype, predicted to be β-thalassemia negative, and found to be different from the allele of the proband at this position, thereby determining the embryo as β - thalassemia is negative.

b) By this method, it can be concluded that the proband and the embryo are the same in the MHC region, so that their HLA type is matched.

According to the β-thalassemia genotype and HLA typing results of the above embryos, the correct embryos, ie, β-thalassemia-negative and HLA-type embryos matched with the proband, are selected for cryopreservation and implantation into the uterus, thereby preventing Pregnancy in sick children. Example 2

First, sample collection and processing:

One couple were carriers of beta-thalassaemia and received IVF-PGD treatment. After an IVF-PGD treatment cycle, 8 blastomere embryos were obtained on the third day, and single-cell biopsy samples of blastomere embryos were taken. The blastomeres after biopsy were cultured until the fifth day of embryogenesis. Embryo vitrification is cryopreserved. Whole genomic amplification of 8 blastomere single cell samples obtained by biopsy, whole genome amplification, using Qiagen's REPLI-g MDA amplification kit, operating strictly according to the kit instructions, amplification products for a new generation Sequencing and other analyses.

At the same time, the peripheral blood of the father and mother in this family was collected and genomic DNA was extracted. After completing the IVF-PGD treatment, the correct embryo is implanted into the uterus. After successful pregnancy, the pregnant women's amniotic fluid can be taken after 14 weeks, or the embryonic cord blood can be collected, DNA extracted, and sequenced to verify the typing results of the embryonic cells.

Eight embryonic single-cell whole-genome amplification product DNA, maternal, maternal peripheral blood DNA, and maternal amniotic fluid or embryonic cord blood DNA were interrupted to a 200 bp fragment using a CovarisTM interrupter, according to the Illumina® HiSeq2000TM sequencer The library is required to build the library, and the library is captured by All in One chip. For details, see the NimbleGen SeqCap EZ chip capture library manual (http://www.nimblegen.com/products/seqcap/ez/index.html). The All in One chip is a self-designed target region capture chip with a target region including a full exon region, a 1M SNP site region, and an MHC gene region. The chip-captured library was quality-controlled by Agilent® Bioanalyzer 2100 and Q-PCR methods. The qualified library was further sequenced by the Illumina® HiSeq2000TM sequencer. The sequencing strategy was Pair End 90 index (ie, bidirectional 90bp index sequencing), and the sequencing depth was 30-50X of the target area. The instrument's parameter settings and operating methods are in accordance with the Illumina® operating manual (available at http://www.illumina.com/support/documentation.ilmn).

Second, sequencing data analysis:

1) Parental genomic DNA, amniotic fluid DNA (validation sample) sequencing data analysis and typing:

a) Filter out sequences of low mass values and joint contamination

b) Use Burrows-Wheeler Aligner (BWA, http /bio-bwa.sourceforge.net/W to capture high shield: ¾|Sequence alignment in human reference genome (HgI9, NCBI release GRCh37) and remove PCR duplication with SAMtools Amplified sequence.

c) Calibrate the alkali matrix mass using the Genome Analysis Toolkit (GATK, http: 〃 www.broadinstitute.org/gatk/) and perform single-base nucleotide polymorphisms and deletions.

2) According to the analysis of 8 single-cell sequencing data of the embryos, the genotype information of the genome initialization of each embryonic cell is given respectively, and the genetic sketch of the genome of each embryo is constructed.

3) Identify any embryos of 8 embryos to be tested, and initialize the haplotypes of the parents according to the genetic sketch of the embryonic cells to be tested and the genotype of the parents.

4) Using the hidden Markov model, reconstruct and determine the parental haplotype according to the genotypes of the remaining 7 embryos.

5) Reconstruct the haplotype of the embryo to be tested according to the parental haplotype by means of the hidden Markov model.

6) Determination of genetic variation information of the embryo to be tested.

Taking chromosome 22 as an example, according to the results of amniotic fluid genotyping, the genotyping accuracy of the embryos to be tested is as follows: Type (female X-father) Total number of sites SNP Call rate accuracy (hybrid) Accuracy (homozygous)

Homozygous X heterozygosity 2183 97.16% 96.68% 99.63%

Hybrid X homozygous 2156 93.18% 98.95% 100%

Hybrid X heterozygous 1154 82.32% 99.31% 97.59%

7) β-thalassemia and HLA type judgment

a) Through the parental haplotype, the haplotype of the embryonic cells to be tested on rs7480526 was successfully reduced, which was predicted to be β-thalassemia negative. b) By using this method, combined with the genetic test results of the infertility, it can be determined whether the haplotypes of the proband and the embryo in the MHC region are consistent, thereby judging whether their HLA types match.

According to the β-thalassemia genotype and HLA typing results of the above embryos, the correct embryos, ie, β-thalassemia-negative and HLA-type embryos matched with the pre-existing animals, are selected for cryopreservation and implantation into the uterus, thereby preventing Pregnancy in sick children. Industrial applicability

The method for determining base region information of a predetermined region in an embryo genome, the system for determining base information of a predetermined region in an embryo genome, and the computer readable medium of the present invention can be effectively applied to analyze a nucleic acid sequence of a predetermined region in an embryo genome. . Although specific embodiments of the invention have been described in detail, those skilled in the art will understand. Various modifications and alterations of those details are possible in light of the teachings of the invention. The full scope of the invention is given by the appended claims and any equivalents thereof.

In the description of the present specification, the description of the terms "one embodiment", "some embodiments", "illustrative embodiment", "example", "specific example", or "some examples", etc. Particular features, structures, materials or features described in the examples or examples are included in at least one embodiment or example of the invention. In the present specification, the schematic representation of the above terms does not necessarily mean the same embodiment or example. Furthermore, the particular features, structures, materials, or characteristics described may be combined in a suitable manner in any one or more embodiments or examples.

Patent Citations (2)

Publication number	Priority date	Publication date	Assignee	Title
CN102127818A *	2010-12-15	2011-07-20	张骥	Method for creating fetus DNA library by utilizing peripheral blood of pregnant woman
CN102317473A *	2009-12-11	2012-01-11	加利福尼亚太平洋生物科学股份有限公司	Shenzhen tci new technology co. , ltd

Family To Family Citations

* Cited by examiner, † Cited by third party

Non-Patent Citations (1)

Title
YU , WEIDONG ET AL.: "Construction of cDNA Library from Rabbit Single Preimplantation Cloned Embryos.", CHINESE JOURNAL OF BIOCHEMISTRY AND MOLECULAR BIOLOGY, vol. 20, no. 2, April 2004 (2004-04-01), pages 162 - 165 *

* Cited by examiner, † Cited by third party

Similar Documents

Publication	Publication Date	Title
US10774380B2	2020-09-15	Methods for multiplex PCR amplification of target loci in a nucleic acid sample
Forsberg et al.	2017	Mosaicism in health and disease—clones picking up speed
US20200251180A1	2020-08-06	Resolving genome fractions using polymorphism counts
Tsang et al.	2017	Integrative single-cell and cell-free plasma RNA transcriptomics elucidates placental cellular dynamics
Kloosterman et al.	2015	Characteristics of de novo structural changes in the human genome
US10174369B2	2019-01-08	Methods for non-invasive prenatal ploidy calling
US20180073080A1	2018-03-15	Methods for non-invasive prenatal ploidy calling
Guo et al.	2014	Three-stage quality control strategies for DNA re-sequencing data
Lo et al.	2016	Limited clinical utility of non-invasive prenatal testing for subchromosomal abnormalities
Gajecka	2016	Unrevealed mosaicism in the next-generation sequencing era
US10597708B2	2020-03-24	Methods for simultaneous amplifications of target loci
Houston et al.	2014	Development and validation of a high density SNP genotyping array for Atlantic salmon (Salmo salar)
JP6388649B2	2018-09-05	Fetal genome analysis of maternal biological samples
Estekl et al.	2015	Concurrent whole-genome haplotyping and copy-number profiling of single cells
Acuna-Hidalgo et al.	2016	New insights into the generation and role of de novo mutations in health and disease
US20200350034A1	2020-11-05	Method for Non-Invasive Prenatal Testing Using Parental Mosaicism Data
Vermeesch et al.	2016	Prenatal and pre-implantation genetic diagnosis
US20190153541A1	2019-05-23	Detecting mutations for cancer screening
US20190050528A1	2019-02-14	Methylation pattern analysis of tissues in a dna mixture
KR102028375B1	2019-10-04	Systems and methods to detect rare mutations and copy number variation
JP6534191B2	2019-06-26	Method for improving the sensitivity of detection in determining copy number variation
Hou et al.	2013	Genome analyses of single human oocytes
AU2017201258B2	2019-04-18	Size-based analysis of fetal dna fraction in maternal plasma
Chen et al.	2016	Second generation noninvasive fetal genome analysis reveals de novo mutations, single-base parental inheritance, and preferred DNA ends
Biesecker et al.	2013	A genomic view of mosaicism and human disease

Priority And Related Applications

Priority Applications (1)

Application	Priority date	Filing date	Title
PCT/CN2013/073375	2013-03-28	2013-03-28	Method, system, and computer readable medium for determining base information of predetermined area in fetal genome

Applications Claiming Priority (3)

Application	Filing date	Title
PCT/CN2013/073375	2013-03-28	Method, system, and computer readable medium for determining base information of predetermined area in fetal genome
CN201380074395.3A	2013-03-28	Method, system, and computer readable medium for determining base information of predetermined area in fetal genome
HK16101707.4A	2016-02-17	Method, system, and computer readable medium for determining base information of predetermined area in fetal genome

Legal Events

Date	Code	Title	Description
2013-03-28	WWE	Wipo information: entry into national phase	**Ref document number:** 201380074395.3 **Country of ref document:** CN
2014-11-19	121	Ep: the epo has been informed by wipo that ep was designated in this application	**Ref document number:** 13879982 **Country of ref document:** EP **Kind code of ref document:** A1
2015-09-28	NENP	Non-entry into the national phase in:	**Ref country code:** DE
2016-05-18	32PN	Ep: public notification in the ep bulletin as address of the addressee cannot be established	**Free format text:** NOTING OF LOSS OF RIGHTS PURSUANT TO RULE 112(1) EPC (EPO FORM 1205A DATED 10/02/2016)
2016-10-05	122	Ep: pct application non-entry in european phase	**Ref document number:** 13879982 **Country of ref document:** EP **Kind code of ref document:** A1

Concepts

machine-extracted

Download Filter table

Name	Image	Sections	Count	Query match
fetal		title,claims,abstract	6	0.000
Embryo, Mammalian		claims,abstract,description	194	0.000
deoxyribonucleic acids		claims,abstract,description	55	0.000
genetic		claims,description	54	0.000
embryonic cell		claims,description	44	0.000
matrix materials		claims,description	38	0.000
embryonic structures		claims,description	25	0.000
analytical methods		claims,description	22	0.000
distribution		claims,description	22	0.000
cells		claims,description	20	0.000
recombination		claims,description	19	0.000
biopsy		claims,description	17	0.000
nucleotide group		claims,description	16	0.000
nucleotides		claims,description	15	0.000
Blastomeres		claims,description	14	0.000
Genetic polymorphism		claims,description	10	0.000
eggs		claims,description	10	0.000

sampling	claims,description	9	0.000
construction	claims,description	7	0.000
solids	claims,description	6	0.000
separation method	claims,description	3	0.000

Show all concepts from the description section

Data provided by IFI CLAIMS Patent Services

About Send Feedback Public Datasets Terms Privacy Policy

(12) 按照专利合作条约所公布的国际申请

(19) 世界知识产权组织
国 际 局

(43) 国际公布日
2014 年 10 月 2 日 (02.10.2014)

WIPO | PCT

(10) 国际公布号
WO 2014/153757 A1

(51) 国际专利分类号:
C12Q 1/68 (2006.01)

(21) 国际申请号: PCT/CN2013/073375

(22) 国际申请日: 2013 年 3 月 28 日 (28.03.2013)

(25) 申请语言: 中文

(26) 公布语言: 中文

(71) 申请人: **深圳华大基因研究院 (BGI SHENZHEN)** [CN/CN]; 中国广东省深圳市盐田区北山工业区综合楼, Guangdong 518083 (CN)。

(72) 发明人: **殷旭阳 (YIN, Xuyang)**; 中国广东省深圳市盐田区北山工业区综合楼, Guangdong 518083 (CN)。 **蒋慧 (JIANG, Hui)**; 中国广东省深圳市盐田区北山工业区综合楼, Guangdong 518083 (CN)。 **陈盛培 (CHEN, Shengpei)**; 中国广东省深圳市盐田区北山工业区综合楼, Guangdong 518083 (CN)。 **龚淳 (GONG, Chun)**; 中国广东省深圳市盐田区北山工业区综合楼, Guangdong 518083 (CN)。 **陈芳 (CHEN, Fang)**; 中国广东省深圳市盐田区北山工业区综合楼, Guangdong 518083 (CN)。 **张春雷 (ZHANG, Chunlei)**; 中国广东省深圳市盐田区北山工业区综合楼, Guangdong 518083 (CN)。 **潘小瑜 (PAN, Xiaoyu)**; 中国广东省深圳市盐田区北山工业区综合楼, Guangdong 518083 (CN)。

(74) 代理人: **北京清亦华知识产权代理事务所（普通合伙） (TSINGYIHUA INTELLECTUAL PROPERTY LLC)**; 中国北京市海淀区清华园清华大学照澜院商业楼 301 室, Beijing 100084 (CN)。

(81) 指定国 (除另有指明，要求每一种可提供的国家保护): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IS, JP, KE, KG, KM, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW。

(84) 指定国 (除另有指明，要求每一种可提供的地区保护): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, SZ, TZ, UG, ZM, ZW), 欧亚 (AM, AZ, BY, KG, KZ, RU, TJ, TM), 欧洲 (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG)。

本国际公布:

— 包括国际检索报告(条约第 21 条(3))。

(54) **Title**: METHOD, SYSTEM, AND COMPUTER READABLE MEDIUM FOR DETERMINING BASE INFORMATION OF PREDETERMINED AREA IN FETAL GENOME

(54) **发明名称** : 确定胚胎基因组中预定区域碱基信息的方法、系统和计算机可读介质

(57) **Abstract**: A method, a system, and a computer readable medium for determining base information of a predetermined area in a fetal genome are provided. The method for determining base information of a predetermined area in a fetal genome comprises the following steps: obtaining a sequencing result of a DNA sample of a genome of an embryo cell, and a sequencing result of a genome sample of an individual related to embryo heredity; constructing a heredity sketch of the embryo based on the sequencing result of the DNA sample of the genome of the embryo cell, to determine an initial genotype of the embryo; determining haplotypes of parents of the embryo based on the sequencing result of the genome sample of the individual related to the embryo heredity; and using the initial genotype of the embryo as an observation sequence according to the hidden markov model, and determining the base information of the predetermined area in the fetal genome based on the haplotypes of the parents of the embryo.

(57) **摘要**: 提供了确定胚胎基因组中预定区域碱基信息的方法、系统和计算机可读介质。其中，确定胚胎基因组中预定区域碱基信息的方法，包括下列步骤：获得胚胎细胞基因组 DNA 样本的测序结果，以及胚胎遗传相关个体基因组样本的测序结果；基于胚胎细胞基因组 DNA 样本的测序结果，构建该胚胎的遗传草图，以便确定胚胎初始基因型；基于胚胎遗传相关个体基因组样本的测序结果，确定胚胎父母的单倍体型；以及根据隐马尔可夫模型，以该胚胎初始基因型作为观察序列，基于该胚胎父母的单倍体型，确定胚胎基因组中预定区域的碱基信息。

WO 2014/153757 A1

确定胚胎基因组中预定区域碱基信息的方法、系统和计算机可读介质

优先权信息

无

技术领域

本发明涉及确定胚胎基因组中预定区域碱基信息的方法、系统和计算机可读介质。

背景技术

遗传性疾病是由于遗传物质发生改变而造成的疾病，具有先天性、家族性、终身性和遗传性的特点。遗传性疾病可分为 3 个大类：单基因遗传病、多基因遗传病及染色体异常。其中单基因病多由于单个致病基因的显性或隐性遗传所致基因功能异常；而多基因遗传病则是由多个基因变化影响所致的疾病，会在一定程度上受到外界环境因素的影响；染色体异常包括数目异常和结构异常，最为多见的是由于第 21 号染色体三体所致的唐氏综合症，患儿表现为先天愚型和肢体形状异常等其他先天性特征。由于目前对遗传性疾病尚无有效的治疗方式，只能针对性地进行支持治疗或者药物缓解，费用昂贵，给社会和家庭带来沉重经济和精神负担。因此，在孩子出生前就对孩子的患病状态进行检测，做好预防工作，以达到优生优育的目的，是十分必要的。例如在体外生殖过程中，在胚胎植入子宫之前，对胚胎的患病状态进行检测，将具有重大意义。对配子或植入到子宫腔之前的胚胎进行遗传学分析，将排除致病因素的正常胚胎植入子宫，从而防止患有遗传病的孩子的妊娠。对具有高风险妊娠遗传缺陷胚胎的夫妇实施植入前胚胎的患病状态的检测，无疑能有效降低患病孩子的出生，并且避免选择性终止妊娠给孕妇及家庭带来的诸多危害。

然而，目前的相关检测手段仍有待改进。

发明内容

本发明是基于发明人的下列发现而完成的：

发明人发现，植入前胚胎中可以活检取样进行检测的细胞数有限。例如培养至第三天的卵裂球仅有 4-8 个细胞，可以进行活检取样的仅 1-2 个细胞。即使在培养至第五天的囊胚，进行外滋养层细胞活检取样也只能取到 3-10 个细胞。由于单个或少数胚胎细胞的 DNA 量有限，难以直接进行全面的遗传变异检测，必需经过全基因组扩增(Whole Genome Amplification, WGA)，达到足够 DNA 量，才能对遗传变异做全面分析。而全基因组扩增常存在偏好性，会产生一定的杂合缺失或等位基因脱扣(Allele Drop Out)等现象，这些都为胚胎细胞的遗传病相关的遗传变异检测带来风险。

因此，很有必要开发能够对扩增引起的等位基因脱扣等错误进行定位和矫正的方法，以提高遗传变异检测的准确性。

本发明旨在至少解决现有技术中存在的技术问题之一。通过本发明方法推算胚胎单

倍体型，可以对发生等位基因脱扣的位点进行定位，并能够进行矫正。而胚胎单倍体型分析方法可以基于家系成员如父母和先证者的遗传信息，或者基于来自同一父母的其他胚胎细胞的遗传信息进行推算。

在本发明的一个方面，本发明提出了一种确定胚胎基因组中预定区域碱基信息的方法。该方法结合系谱推断和统计算法来进行胚胎单体型的确定，通过家系中相关个体的基因型，即追溯染色体片段的传递，结合统计算法来推断单体型状态。根据本发明的实施例，该方法包括下列步骤：获得胚胎细胞基因组 DNA 样本的测序结果，以及胚胎遗传相关个体基因组样本的测序结果；基于胚胎细胞基因组 DNA 样本的测序结果，构建所述胚胎的遗传草图，以便确定胚胎初始基因型；基于所述胚胎遗传相关个体基因组样本的测序结果，确定胚胎父母的单倍体型；以及根据隐马尔可夫模型，以所述胚胎初始基因型作为观察序列，基于所述胚胎父母的单倍体型，确定胚胎基因组中预定区域的碱基信息。

需要说明的是，子代的基因组形成过程，相当于亲代基因组的一次随机重组（即连锁互换单倍体型重组，以及配子的随机组合）。将胚胎的单倍型作为隐含状态（hidden states），将胚胎单细胞全基因组扩增后的测序数据作为观察序列（observations），利用贝叶斯算法借助先验数据推算出状态转移概率（transition probabilities），构建观察序列概率分布（observation symbol probabilities）和初始状态概率分布（initial state distribution），然后通过惠特比算法（Viterbi algorithm）可以推断出最可能的胚胎单体型组合。因而，根据本发明的实施例，借助隐马尔可夫模型，例如可以通过利用惠特比算法（Viterbi algorithm），参考胚胎遗传相关个体的遗传信息，可以确定胚胎基因组中特定区域的核酸序列，由此，可以有效地对胚胎基因组的遗传信息进行植入前检测。对于在胚胎细胞全基因组扩增 DNA 测序结果中覆盖程度较低，存在杂合缺失或等位基因脱扣(Allele Drop Out)的位点或 DNA 序列，可以通过本方法准确地推断得到。因此通过本方法，可以对杂合缺失或等位基因脱扣(Allele Drop Out)的位点或序列进行校正，使检测结果更加准确可靠。

在本发明的又一方面，本发明提出了一种用于确定胚胎基因组中预定区域碱基信息的系统。根据本发明的实施例，该系统包括：文库构建装置，所述文库构建装置适于针对胚胎基因组 DNA 样本以及胚胎遗传相关个体的基因组 DNA 样本，分别构建测序文库；测序装置，所述测序装置与所述文库构建装置相连，并且适于对所述测序文库进行测序，以便获得胚胎以及胚胎遗传相关个体的测序结果；以及分析装置，所述分析装置与所述测序装置相连，并且适于：基于胚胎基因组 DNA 样本的测序结果，构建所述胚胎的遗传草图，以便确定胚胎初始基因型；基于所述胚胎遗传相关个体基因组样本的测序结果，确定胚胎父母的单倍体型；以及根据隐马尔可夫模型，以所述胚胎初始基因型作为观察序列，基于所述胚胎父母的单倍体型，确定胚胎基因组中预定区域的碱基信息。根据本发明的实施例，还可以进一步包括胚胎活检装置，所述胚胎活检装置适于针对体外培养的胚胎进行活检，对胚胎细胞进行取样。利用该系统，能够有效地实施前面所述的确定胚胎基因组中预定区域碱基信息的方法，可以借助隐马尔可夫模型，例如可以通过利用惠特比算法（Viterbi algorithm），参考胚胎遗传相关个体的遗传信息，可以确定胚胎基因组中特定区域的核酸序

列，由此，可以有效地对胚胎基因组的遗传信息进行植入前检测，从而可以有效地对胚胎基因组的遗传信息进行确定。

在本发明的另一方面，本发明还提出了一种计算机可读介质。根据本发明的实施例，基于胚胎细胞基因组DNA样本的测序结果，构建胚胎的遗传草图，以便确定胚胎初始基因型；基于所述胚胎遗传相关个体基因组样本的测序结果，确定胚胎父母的单倍体型；以及以根据隐马尔可夫模型，所述胚胎初始基因型作为观察序列，基于所述胚胎父母的单倍体型，确定胚胎基因组中预定区域的碱基信息。利用本发明的计算机可读介质，能够有效地被处理器执行其存储的指令，以便借助隐马尔可夫模型，例如可以通过利用惠特比算法（Viterbi algorithm），基于胚胎细胞的测序结果，参考胚胎遗传相关个体的遗传信息，可以确定胚胎基因组中特定区域的核酸序列，由此，可以有效地对胚胎基因组的遗传信息进行植入前检测。

本发明的附加方面和优点将在下面的描述中部分给出，部分将从下面的描述中变得明显，或通过本发明的实践了解到。

附图说明

本发明的上述和/或附加的方面和优点从结合下面附图对实施例的描述中将变得明显和容易理解，其中：

图1为根据本发明一个实施例的利用隐马尔可夫模型进行分析的流程示意图；以及

图2为根据本发明的一个实施例的用于确定胚胎基因组中预定区域核酸序列的系统的结构示意图。

发明详细描述

下面详细描述本发明的实施例，所述实施例的示例在附图中示出，其中自始至终相同或类似的标号表示相同或类似的元件或具有相同或类似功能的元件。下面通过参考附图描述的实施例是示例性的，仅用于解释本发明，而不能理解为对本发明的限制。

需要说明的是，术语“第一”、“第二”仅用于描述目的，而不能理解为指示或暗示相对重要性或者隐含指明所指示的技术特征的数量。由此，限定有“第一”、“第二”的特征可以明示或者隐含地包括一个或者更多个该特征。进一步地，在本发明的描述中，除非另有说明，“多个”的含义是两个或两个以上。

确定胚胎基因组中预定区域碱基信息的方法

在本发明的第一方面，本发明提出了一种确定胚胎基因组中预定区域碱基信息的方法。根据本发明的实施例，该方法包括下列步骤：

首先，获得胚胎细胞基因组DNA样本的测序结果，以及胚胎遗传相关个体基因组样本的测序结果。其中，在本文中所使用的术语“胚胎遗传相关个体”指的是在遗传意义上，与胚胎之间具有亲缘关系的个体，例如根据本发明的实施例，可以采用的“胚胎遗传相关个体”为胚胎的亲代例如父母、祖父母、外祖父母以及胚胎父母的其他子女。这里所说的

胚胎父母的其他子女应做广义理解，既可以是已经出生的子女，也可以是尚未出生的子女（受精卵或胚胎），也可以是死亡的胚胎，也可以是体外培养的胚胎或受精卵，只要与待检测胚胎共有相同的父母即可。

根据本发明的实施例，胚胎细胞基因组 DNA 样本的来源并不受特别限制。根据本发明的实施例，具体地，可以针对胚胎进行活检取样，获得胚胎细胞，并对胚胎细胞样本进行全基因组扩增(WGA)，以便获得胚胎细胞基因组 DNA。在本文中所采用的术语“胚胎活检”是指利用显微技术，从胚胎中分离部分胚胎细胞，或从受精卵/配子中分离部分细胞的技术。其中，根据本发明的实施例，该胚胎细胞可来自于卵裂球、囊胚滋养层、受精卵和配子中的任一种，可以是单个细胞，也可以是 2-10 个细胞。此外，也可以采用任何含有胚胎核酸的孕妇样本进行全基因组扩增，以便获得胚胎基因组 DNA 样本。由此，可以在不影响胚胎正常发育的前提下，对胚胎的基因组进行有效监测。其中，“全基因组扩增(WGA)”主要包括多重链置换扩增(Multiple Displacement Amplification, MDA)和基于 PCR 的 WGA 技术，可以采用自主开发的 WGA 扩增流程，也可以采用商业化的试剂盒如 Qiagen 公司的 REPLI-g 系列的试剂盒、Sigma-Aldrich 公司的 GenomePlex WGA 试剂盒(WGA4)、Rubicon Genomics 公司的 PicoPlex WGA 试剂盒、GE-Healthcare 公司的 illustra Genomiphi WGA 试剂盒等。在扩增 DNA 样本之后，针对胚胎细胞的扩增 DNA 样本以及胚胎遗传相关个体的基因组 DNA 样本，分别构建测序文库。

关于针对核酸样本，构建测序文库的方法和流程，本领域技术人员可以根据不同的测序技术进行适当选择，关于流程的细节，可以参见测序仪器的厂商例如 Illumina 公司所提供的规程，例如参见 Illumina 公司 Multiplexing Sample Preparation Guide （Part#1005361; Feb 2010）或 Paired-End Sample Prep Guide（Part#1005063; Feb 2010），通过参照将其并入本文。根据本发明的实施例，从生物样本提取核酸样本的方法和设备，也不受特别限制，可以采用商品化的核酸提取试剂盒进行。

在构建测序文库后，将测序文库应用于测序仪器，对测序文库进行测序，并获得相应的测序结果，该测序结果是由多个测序数据构成的。根据本发明的实施例，可以用于进行测序的方法和设备并不受特别限制，包括但不限于双脱氧链终止法；优选高通量的测序方法，由此，能够利用这些测序装置的高通量、深度测序的特点，进一步提高测序效率。从而，能够提高后续对测序数据进行分析，尤其是统计检验分析时的精确性和准确度。所述高通量的测序方法包括但不限于第二代测序技术或者是单分子测序技术。所述第二代测序平台（Metzker ML. Sequencing technologies-the next generation. Nat Rev Genet. 2010 Jan;11(1):31-46）包括但不限于 Illumina-Solexa（GA™,HiSeq2000™, Miseq 等）、ABI-SOLiD、Life Technologies-Ion Torrent/Proton 和 Roche-454（焦磷酸测序）测序平台；单分子测序平台（技术）包括但不限于 Helicos 公司的真实单分子测序技术（True Single Molecule DNA sequencing），Pacific Biosciences 公司单分子实时测序（single molecule real-time (SMRT™)），以及 Oxford Nanopore Technologies 公司的纳米孔测序技术等（Rusk, Nicole (2009-04-01). Cheap Third-Generation Sequencing. Nature Methods 6 (4): 244-245）。随着测序

技术的不断进化，本领域技术人员能够理解的是还可以采用其他的测序方法和装置进行全基因组测序。根据本发明的具体示例，可以利用选自 Illumina-Solexa、ABI-SOLiD、Life Technologies-Ion Torrent/Proton、Roche-454 和单分子测序装置的至少一种对所述全基因组测序文库进行测序。

根据本发明的实施例，在得到测序结果之后，可以基于胚胎细胞基因组 DNA 样本的测序结果，构建所述胚胎的遗传草图，以便确定胚胎初始基因型。基于所述胚胎遗传相关个体基因组样本的测序结果，确定胚胎父母的单倍体型。根据本发明的实施例，通过将胚胎细胞基因组 DNA 样本的测序结果与参考序列进行比对，构建所述胚胎的遗传草图。通过将所述胚胎遗传相关个体基因组样本的测序结果与参考序列进行比对，确定所述胚胎遗传相关个体的基因型；以及基于所述胚胎遗传相关个体的基因型，确定所述胚胎父母的单倍体型。根据本发明的实施例，可以采用已知的人类参考基因组作为参考序列。例如，根据本发明的实施例，采用的人类参考基因组为 NCBI 36.3，HG18。另外，根据本发明的实施例，进行比对的方法并不受特别限制。根据本发明的具体实施例，可以采用 SOAP 进行比对。

最后，以所述胚胎初始基因型作为观察序列，基于所述胚胎父母的单倍体型，根据隐马尔可夫模型，确定胚胎基因组中预定区域的碱基信息。

在本文中所使用的术语“预定区域”应作广义理解，是指任何包含可能发生预定事件位点的核酸分子的区域。对于 SNP 分析而言，可以是指包含 SNP 位点的区域。对于分析染色体非整倍性，则预定区域指的是所要分析的染色体的全长或者部分，即选择所有来自该染色体的测序数据。从测序结果中选择来自相应区域的测序数据的方法可以不受特别限制。根据本发明的实施例，可以通过将所得到的所有测序数据与已知的核酸参考序列进行比对，从而得到来自于预定区域的测序数据。另外，根据本发明的实施例，预定区域也可以是基因组上不连续的多个分散点。根据本发明的实施例，可以使用的参考序列的类型并不受特别限制，可以为任何含有感兴趣区域的已知序列。

根据本发明的实施例，可以通过基于所述胚胎细胞的测序结果，结合胚胎遗传相关个体的遗传信息，根据隐马尔可夫模型，确定所述预定区域的碱基信息。根据本发明的实施例，可以借助隐马尔可夫模型，利用惠特比算法（Viterbi algorithm），确定胚胎基因组中特定区域的碱基信息。由此，可以有效地对胚胎基因组的遗传信息进行植入前检测。

下面参考图 1，对借助隐马尔可夫模型利用惠特比算法进行分析的原理进行详细描述：

如前所述，在本文中所使用的术语“胚胎遗传相关个体”指的是在遗传意义上，与胚胎之间具有亲缘关系的个体，例如根据本发明的实施例，可以采用的“胚胎遗传相关个体”为胚胎的亲代例如父母、祖父母、外祖父母以及胚胎父母的其他子女。这里所说的胚胎父母的其他子女应做广义理解，既可以是已经出生的子女，也可以是尚未出生的子女（胚胎或受精卵），也可以是死亡的胚胎，也可以是体外培养的胚胎或受精卵，只要与待检测胚胎共有相同的父母即可。

由此，子代的基因组形成过程，相当于亲代基因组的一次随机重组（即连锁互换单倍体型重组，以及配子的随机组合）。将胚胎的单倍型作为隐含状态（hidden states），将胚胎

单细胞全基因组扩增后的测序数据作为观察序列（observations），借助先验数据推算出状态转移概率（transition probabilities），构建确定观察序列概率分布（observation symbol probabilities）和初始状态概率分布（initial state distribution），然后通过惠特比算法（Viterbi algorithm）可以推断出最可能的胚胎单体型组合。因而，根据本发明的实施例，借助隐马尔可夫模型，例如可以通过利用惠特比算法（Viterbi algorithm），参考胚胎遗传相关个体的遗传信息，可以确定胚胎基因组中特定区域的核酸序列，由此，可以有效地对胚胎基因组的遗传信息进行植入前检测。

详细分析步骤如下：

标记：

I. 需要检测的位点数为 N。

II. 观测序列；

III. 隐藏状态集为：$S=\{0,1\}$，定义了父母哪条染色体遗传给了子代。0 代表遗传给了先证者的那条染色体，1 代表没有遗传给先证者的那条染色体。

IV. 观测状态集为：$V=\{0,1\}$，1 代表遗传给先证者的染色体与胚胎遗传草图基因型一致，0 代表不一致。

第一步，构建初始状态概率分布向量，单倍体重组转移矩阵以及观察序列概率矩阵

I. 初始状态概率分布记为 $\pi=\{\pi_i\},(i\in S,\pi_i=0.5)$，

在没有参考数据的情况下，可以假设每种隐藏状态出现的可能性相等。

II. 记单倍体重组转移矩阵记为 $A=\{a_{ij}\},(i,j\in S)$，即隐藏状态转移的概率。

$$a_{ij}=\begin{cases}1-Nr/Np & i=j\\ Nr/Np & i\neq j\end{cases},$$

Nr、Np 分别表示期望重组数和单核苷酸多态性位点数，根据先验数据 Nr 可取 20-40 之间的自然数，用来计算隐藏状态的转移概率，即计算每个碱基组合的单体型发生重组的概率。

III. 记观察序列概率矩阵为 $B=\{b_i(k)\},(i\in S,k\in V)$，这里的概率分布 $b_i(k)$ 用子代一定是纯合（Must-hom）或一定是杂合(Must-het)的位点来估算。

$$b_i(0)=\frac{\#sites(L>0,\text{Must-hom.})}{\#sites(L>0,\text{Must-het. or Must-hom.})},\quad b_i(1)=1-b_i(0)。$$

第二步，构建局部概率矩阵和逆向指针

定义 局部概率：$\delta_t(j)=\max\left[\delta_{t-1}(i)\cdot a_{ij}\right]\cdot b_j(K)\quad t\in\{1...N\}$。

定义 逆向指针 ：$\Psi_i(j)=\arg\max \delta_{t-1}(i)\cdot a_{ij} \quad t\in\{1...N\}$。

第三步，确定最终状态状态，并回溯最优路径

确定最终状态，$q_N^* = \arg\max_{q_N\in S}\delta_N(q_N)$。

回溯最优路径，即最可能胚胎单基因型 $q_t^* = \Psi_{t+1}(q^*_{t+1})$ $(t=1,2,3,...,N-1)$。

第四步，输出结果

即，根据本发明的实施例，根据隐马尔可夫模型，确定胚胎基因组中预定区域的碱基信息进一步包括:

构建初始状态概率分布向量、隐藏状态转移的概率矩阵以及观察序列概率矩阵;

利用惠特比算法确定最终状态并回溯最优路径，以便确定胚胎基因组中预定区域的碱基信息。

根据具体的实施例，所述隐马尔可夫模型采用下列参数:

初始状态概率分布为 $\pi=\{\pi_i\},(i\in S,\pi_i=0.5)$,

隐藏状态转移的概率矩阵为 $A=\{a_{ij}\},(i,j\in S)$，其中，$a_{ij}=\begin{cases}1-Nr/Np & i=j\\ Nr/Np & i\neq j\end{cases}$

Nr、Np 分别表示期望重组数和单核苷酸多态性位点数，Nr 为自然数，取值范围 20-40，

观察序列概率矩阵为 $B=\{b_i(k)\},(i\in S,k\in V)$，其中，$b_i(0)=\frac{\#sites(L>0,\text{Must-hom.})}{\#sites(L>0,\text{Must-het. or Must-hom.})}$, $b_i(1)=1-b_i(0)$,

#sites (L>0,Must-hom.)为子代一定为纯合的位点的数目，#sites (L>0,Must-hom. or Must-het.)为子代一定为纯合的位点的数目以及子代一定为杂合的位点的数目的总和;

局部概率为 $\delta_t(j)=\max\left[\delta_{t-1}(i)\cdot a_{ij}\right]\cdot b_j(K) \quad t\in\{1...N\}$,

逆向指针为 $\Psi_i(j)=\arg\max \delta_{t-1}(i)\cdot a_{ij} \quad t\in\{1...N\}$,

依惠特比算法经递归得最终状态为 $q_N^* = \arg\max_{q_N\in S}\delta_N(q_N)$,

回溯最优路径，确定最可能的胚胎预定区域的碱基信息为 $q_t^* = \Psi_{t1}(q^*_{t+1})$ $(t=1,2,3,...,N-1)$。

其中，这里所使用的术语“局部概率$\delta_i(q_i)$”和“逆向指针$\Psi_i(q_i)$”都是沿用 Viterbi 算法的经典定义。关于该参数的定义的详细描述，可以参见 Lawrence R. Rabiner, PROCEEDINGS OF THE IEEE， Vol.77，No.2，1989 年 2 月，通过参照将其全文并入本文。

由此，能够有效地对胚胎基因组的序列进行分析。相比其他已有的植入前检测技术方法，本方法有以下技术优势，主要体现在准确性和可获得的遗传信息量上:

1) 根据本发明实施例，针对胚胎的父源性位点和母源性位点，均可以很好地检测，检测准确率可高达 95%以上，且可以检测多种变异类型，扩大了疾病检测的范围。

2) 根据本发明实施例，对于一些在胚胎单细胞或微量细胞全基因组扩增 DNA 测序结果中覆盖程度较低，存在杂合缺失或等位基因脱扣(Allele Drop Out)的位点或 DNA 序列，可以通过本方法准确地推断得到。因此通过本方法，可以对杂合缺失或等位基因脱扣(Allele Drop Out)的位点或序列进行校正，使检测结果更加准确可靠。

3) 根据本发明实施例，可进行遗传疾病作图，对于一些连锁相关疾病，可通过其他位点的信息直接推断出来，一次可获得的信息量大，对临床检测更加具有指导意义。

另外，根据本发明实施例，本发明的确定胚胎基因组中预定区域碱基信息的方法，不仅限于 SNP 或者 STR 等某一种遗传多态性位点，对所有的遗传多态性位点均可适用，且可以多种位点同时使用，以便互相验证。

用于确定胚胎基因组中预定区域碱基信息的系统

在本发明的又一方面，本发明提出了一种用于确定胚胎基因组中预定区域核酸序列的系统。根据本发明的实施例，参考图 2，该系统 1000 可以包括: 文库构建装置 100、测序装置 200 以及分析装置 400。

根据本发明的实施例，文库构建装置 100 适于针对胚胎基因组 DNA 样本以及胚胎遗传相关个体的基因组 DNA 样本，分别构建测序文库。根据本发明的实施例，测序装置 200 与文库构建装置 100 相连，并且适于对所述测序文库进行测序，以便获得胚胎以及胚胎遗传相关个体的测序结果。根据本发明的实施例，还可以进一步包括 DNA 样本分离装置和 DNA 扩增装置（图中未示出），该 DNA 样本分离装置适于针对胚胎进行活检取样，获得胚胎细胞。胚胎细胞可来自于卵裂球、囊胚滋养层、受精卵和配子中的任一种，可以是单个细胞，也可以是微量细胞例如 2-10 个细胞，也可以采用任何含有胚胎核酸的孕妇样本。该 DNA 扩增装置适于针对活检取样获得的胚胎细胞进行全基因组扩增，以获得足够量 DNA 用于后续检测分析。由此，可以在不影响胚胎正常发育的前提下，对胚胎的基因组进行有效监测。关于全基因组扩增的方法和流程，主要包括多重链置换扩增(Multiple Displacement Amplification, MDA)和基于 PCR 的 WGA 技术，可以采用自主开发的 WGA 扩增流程，也可以采用商业化的试剂盒如 Qiagen 公司的 REPLI-g 系列的试剂盒、Sigma-Aldrich 公司的 GenomePlex WGA 试剂盒(WGA4)、Rubicon Genomics 公司的 PicoPlex WGA 试剂盒、GE-Healthcare 公司的 illustra Genomiphi WGA 试剂盒等等。

关于针对核酸样本，构建测序文库的方法和流程，本领域技术人员可以根据不同的测序技术进行适当选择，关于流程的细节，可以参见测序仪器的厂商例如 Illumina 公司所提供的规程，例如参见 Illumina 公司 Multiplexing Sample Preparation Guide （Part#1005361; Feb 2010）或 Paired-End SamplePrep Guide（Part#1005063; Feb 2010），通过参照将其并入本文。根据本发明的实施例，从生物样本提取核酸样本的方法和设备，也不受特别限制，可以采用商品化的核酸提取试剂盒进行。根据本发明的实施例，可以用于进行测序的方法和设备并不受特别限制，包括但不限于双脱氧链终止法；优选高通量的测序方法，由此，能够利用这些测序装置的高通量、深度测序的特点，进一步提高测序效率。从而，能够提高后续对测序数据进行分析，尤其是统计检验分析时的精确性和准确度。所述高通量的测序方法包括但不限于第二代测序技术或者是单分子测序技术。所述第二代测序平台（Metzker ML. Sequencing technologies-the next generation. Nat Rev Genet. 2010 Jan;11(1):31-46）包括但不限于 Illumina-Solexa（GA™、HiSeq™、Miseq 等）、ABI-SOLiD、Life Technologies-Ion Torrent/Proton 和 Roche-454（焦磷酸测序）测序平台；单分子测序平台（技术）包括但不限于 Helicos 公司的真实单分子测序技术（True Single Molecule DNA sequencing）， Pacific Biosciences 公司单分子实时测序（single molecule real-time (SMRT™)），以及 Oxford Nanopore Technologies 公司的纳米孔测序技术等（Rusk, Nicole (2009-04-01). Cheap Third-Generation Sequencing. Nature Methods 6 (4): 244-245）。随着测序技术的不断进化，本领域技术人员能够理解的是还可以采用其他的测序方法和装置进行全基因组测序。根据本发明的具体示例，可以利用选自 Illumina-Solexa、ABI-SOLiD、Life Technologies-Ion Torrent/Proton、Roche-454 和单分子测序装置的至少一种对所述全基因组测序文库进行测序。

根据本发明的实施例，还可以包括比对装置 300。根据本发明的实施例，比对装置 300 与测序装置 200 相连，并且适于将所得到的测序结果与参考序列进行比对，以便通过将胚胎细胞基因组 DNA 样本的测序结果与参考序列进行比对，构建所述胚胎的遗传草图；通过将所述胚胎遗传相关个体基因组样本的测序结果与参考序列进行比对，确定所述胚胎遗传相关个体的基因型；以及基于所述胚胎遗传相关个体的基因型，确定所述胚胎父母的单倍体型。根据本发明的实施例，可以使用的参考序列的类型并不受特别限制，可以为任何含有感兴趣区域的已知序列。根据本发明的实施例，可以采用已知的人类参考基因组作为参考序列。例如，根据本发明的实施例，采用的人类参考基因组为 NCBI 36.3，HG18。另外，根据本发明的实施例，进行比对的方法并不受特别限制。根据本发明的具体实施例，可以采用 SOAP 进行比对。

根据本发明的实施例，分析装置 400 适于：基于胚胎细胞基因组 DNA 样本的测序结果，构建所述胚胎的遗传草图，以便确定胚胎初始基因型；基于所述胚胎遗传相关个体基因组样本的测序结果，确定胚胎父母的单倍体型；以及根据隐马尔可夫模型，以所述胚胎初始基因型作为观察序列，基于所述胚胎父母的单倍体型，确定胚胎基因组中预定区域的碱基信息。

根据本发明的实施例，根据隐马尔可夫模型，确定胚胎基因组中预定区域的碱基信息

进一步包括:

构建初始状态概率分布向量、隐藏状态转移的概率矩阵以及观察序列概率矩阵;

利用惠特比算法确定最终状态并回溯最优路径，以便确定胚胎基因组中预定区域的碱基信息。

根据具体的实施例，所述隐马尔可夫模型采用下列参数:

初始状态概率分布为 $\pi=\{\pi_i\},(i\in S,\pi_i=0.5)$,

隐藏状态转移的概率矩阵为 $A=\{a_{ij}\},(i,j\in S)$，其中，$a_{ij}=\begin{cases}1-Nr/Np & i=j\\ Nr/Np & i\neq j\end{cases}$

Nr, Np 分别表示期望重组数和单核苷酸多态性位点数，Nr 可取 20-40 之间的自然数，

观察序列概率矩阵为 $B=\{b_i(k)\},(i\in S,k\in V)$，其中，

$b_i(0)=\frac{\#sites(L>0,\text{Must-hom.})}{\#sites(L>0,\text{Must-het. or Must-hom.})}$, $b_i(1)=1-b_i(0)$,

#sites (L>0,Must-hom.)为子代一定为纯合的位点的数目，#sites (L>0,Must-hom. or Must-het.)为子代一定为纯合的位点的数目以及子代一定为杂合的位点的数目的总和;

局部概率为 $\delta_t(j)=\max\left[\delta_{t-1}(i)\cdot a_{ij}\right]\cdot b_j(K)\quad t\in\{1...N\}$,

逆向指针为 $\Psi_i(j)=\arg\max\delta_{t-1}(i)\cdot a_{ij}\quad t\in\{1...N\}$,

递归得最终状态为 $q_N^*=\arg\max_{q_N\subset S}\delta_N(q_N)$,

回溯最优路径，确定最可能的胚胎预定区域的碱基信息为 $q_t^*=\Psi_{t+1}(q^*_{t+1})$ $(t=1,2,3,...,N-1)$。

这里所使用的术语局部概率 $\delta_i(q_i)$和逆向指针 $\Psi_i(q_i)$都是沿用 Viterbi 算法的经典定义。关于该参数的定义的详细描述，可以参见 Lawrence R. Rabiner, PROCEEDINGS OF THE IEEE， Vol.77，No.2，1989 年 2 月，通过参照将其全文并入本文。

关于数据分析部分，前面已经进行了详细描述，也当然地适用于确定胚胎基因组中预定区域核酸序列的系统。不再赘述。

由此，利用该系统，能够有效地实施前面所述的确定胚胎基因组中预定区域核酸序列的方法，可以借助隐马尔可夫模型，通过例如惠特比算法（Viterbi algorithm）解码，确定胚胎基因组中特定区域的碱基信息，由此，可以有效地对胚胎基因组的遗传信息进行植入前检测。

此外，根据本发明的实施例，预定区域为已知存在遗传多态性的位点，而遗传多态性为选自单核苷酸多态性和 STR 的至少一种。

在本文中所述的术语“相连”应作广义理解，既可以是直接相连，也可以是间接相连，只要能够实现上述功能上的衔接即可。

需要说明的是，本领域技术人员能够理解，在前面所描述的确定胚胎基因组中预定区域核酸序列的方法的特征和优点也适合于确定胚胎基因组中预定区域核酸序列的系统，为描述方便，不再详述。

计算机可读介质

在本发明的又一方面，本发明提出了一种计算机可读介质。根据本发明的实施例，所述计算机可读介质上存储有指令，所述指令适于被处理器执行以便:

基于胚胎细胞基因组 DNA 样本的测序结果，构建胚胎的遗传草图，以便确定胚胎初始基因型;

基于该胚胎遗传相关个体基因组样本的测序结果，确定胚胎父母的单倍体型；以及

根据隐马尔可夫模型，以该胚胎初始基因型作为观察序列，基于该胚胎父母的单倍体型，确定胚胎基因组中预定区域的碱基信息。

由此，利用该计算机可读介质，能够有效地实施前面所述的方法，从而可以通过例如惠特比算法（Viterbi algorithm），借助隐马尔可夫模型，确定胚胎基因组中特定区域的碱基信息，由此，可以有效地对胚胎基因组的遗传信息进行植入前检测。

根据本发明的实施例，根据隐马尔可夫模型，确定胚胎基因组中预定区域的碱基信息进一步包括:

构建初始状态概率分布向量、隐藏状态转移的概率矩阵以及观察序列概率矩阵;

利用惠特比算法确定最终状态并回溯最优路径，以便确定胚胎基因组中预定区域的碱基信息。

根据具体的实施例，所述隐马尔可夫模型采用下列参数:

初始状态概率分布为 $\pi=\{\pi_i\},(i\in S,\pi_i=0.5)$,

隐藏状态转移的概率矩阵为 $A=\{a_{ij}\},(i,j\in S)$，其中，$a_{ij}=\begin{cases}1-Nr/Np & i=j\\ Nr/Np & i\neq j\end{cases}$

Nr、Np 分别表示期望重组数和单核苷酸多态性位点数，Nr 可取 20-40 之间的自然数，

观察序列概率矩阵为 $B=\{b_i(k)\},(i\in S,k\in V)$，其中，

$b_i(0)=\frac{\#sites(L>0,\text{Must-hom.})}{\#sites(L>0,\text{Must-het. or Must-hom.})}$, $b_i(1)=1-b_i(0)$,

#sites (L>0,Must-hom.)为子代一定为纯合的位点的数目，#sites (L>0,Must-hom. or Must-het.)为子代一定为纯合的位点的数目以及子代一定为杂合的位点的数目的总和;

局部概率为 $\delta_t(j)=\max\left[\delta_{t-1}(i)\cdot a_{ij}\right]\cdot b_j(K) \quad t\in\{1...N\}$

逆向指针为 $\Psi_i(j)=\arg\max\delta_{t-1}(i)\cdot a_{ij} \quad t\in\{1...N\}$，

经递归得最终状态为 $q_N^*=\arg\max_{q_N\subset S}\delta_N(q_N)$，

回溯最优路径，确定最可能的胚胎预定区域的碱基信息为 $q_t^*=\Psi_{t+1}(q^*_{t+1})$ $(t=1,2,3,...,N-1)$。

这里所使用的术语局部概率$\delta_i(q_i)$和逆向指针$\Psi_i(q_i)$都是沿用 Viterbi 算法的经典定义。关于该参数的定义的详细描述，可以参见 Lawrence R. Rabiner, PROCEEDINGS OF THE IEEE， Vol.77，No.2，1989 年 2 月，通过参照将其全文并入本文。

关于数据分析部分，前面已经进行了详细描述，也当然地适用于确定胚胎基因组中预定区域核酸序列的系统。不再赘述。

由此，利用该系统，能够有效地实施前面所述的确定胚胎基因组中预定区域核酸序列的方法，可以通过例如惠特比算法（Viterbi algorithm），借助隐马尔可夫模型，确定胚胎基因组中特定区域的碱基信息，由此，可以有效地对胚胎基因组的遗传信息进行植入前检测。

此外，根据本发明的实施例，预定区域为已知存在遗传多态性的位点，而遗传多态性为选自单核苷酸多态性和 STR 的至少一种。

就本说明书而言，“计算机可读介质”可以是任何可以包含、存储、通信、传播或传输程序以供指令执行系统、装置或设备或结合这些指令执行系统、装置或设备而使用的装置。计算机可读介质的更具体的示例（非穷尽性列表）包括以下：具有一个或多个布线的电连接部（电子装置），便携式计算机盘盒（磁装置），随机存取存储器（RAM），只读存储器（ROM），可擦除可编辑只读存储器（EPROM 或闪速存储器），光纤装置，以及便携式光盘只读存储器（CDROM）。另外，计算机可读介质甚至可以是可在其上打印所述程序的纸或其他合适的介质，因为例如可以通过对纸或其他介质进行光学扫描，接着进行编辑、解译或必要时以其他合适方式进行处理来以电子方式获得所述程序，然后将其存储在计算机存储器中。

应当理解，本发明的各部分可以用硬件、软件、固件或它们的组合来实现。在上述实施方式中，多个步骤或方法可以用存储在存储器中且由合适的指令执行系统执行的软件或固件来实现。例如，如果用硬件来实现，和在另一实施方式中一样，可用本领域公知的下列技术中的任一项或他们的组合来实现：具有用于对数据信号实现逻辑功能的逻辑门电路的离散逻辑电路，具有合适的组合逻辑门电路的专用集成电路，可编程门阵列（PGA），现场可编程门阵列（FPGA）等。

本技术领域的普通技术人员可以理解实现上述实施例方法携带的全部或部分步骤是可以通过程序来指令相关的硬件完成，所述的程序可以存储于一种计算机可读存储介质中，

该程序在执行时，包括方法实施例的步骤之一或其组合。

此外，在本发明各个实施例中的各功能单元可以集成在一个处理模块中，也可以是各个单元单独物理存在，也可以两个或两个以上单元集成在一个模块中。上述集成的模块既可以采用硬件的形式实现，也可以采用软件功能模块的形式实现。所述集成的模块如果以软件功能模块的形式实现并作为独立的产品销售或使用时，也可以存储在一个计算机可读取存储介质中。

下面将结合实施例对本发明的方案进行解释。本领域技术人员将会理解，下面的实施例仅用于说明本发明，而不应视为限定本发明的范围。实施例中未注明具体技术或条件的，按照本领域内的文献所描述的技术或条件（例如参考 J.萨姆布鲁克等著，黄培堂等译的《分子克隆实验指南》，第三版，科学出版社）或者按照产品说明书进行。所用试剂或仪器未注明生产厂商者，均为可以通过市购获得的常规产品，例如可以采购自 Illumina 公司。

一般方法:

参考图 1，本发明实施例的主要步骤包括:

1) 获得单细胞或微量细胞样本，第三天卵裂球细胞或第五天囊胚滋养层细胞活检取样，或受精卵取样，或其他单细胞或微量细胞取样。
2) 单细胞或微量细胞样本进行全基因组扩增，扩增产物 DNA 根据新一代测序平台要求构建文库，并进行测序。
3) 单细胞或微量细胞测序数据经过过滤后与人类基因组参考序列进行比对。
4) 根据比对结果进行胚胎细胞单核苷酸多态性和缺失重复等遗传变异的初步检测。
5) 构建胚胎基因组的遗传草图，对胚胎基因型进行初始化。
6) 收集胚胎遗传相关个体，如家系成员如父母和先证者，和/或祖父母、外祖父母等或者同一对父母的其它胚胎样本，提取基因组 DNA，根据新一代测序平台要求构建文库，并进行测序。
7) 胚胎遗传相关个体 DNA 测序数据进行过滤，并与人类基因组参考序列比对。
8) 根据比对结果，分析确定先证者、父母等遗传相关个体的基因型。
9) 利用先证者、父母等遗传相关个体的基因型，推断父母双亲的单倍体型。
10) 借助隐马尔科夫模型解码方式，用父母单倍体型结果推断胚胎单体型。
11) 胚胎遗传变异的最终确定。

参考图 1，本发明实施例中信息分析部分用到的隐马尔科夫模型以及详细分析步骤如下:

标记:

1、需要检测的位点数为 N。

2、观测序列:

3、隐藏状态集为: $S=\{0,1\}$，定义了父母哪条染色体遗传给了子代。0 代表遗传给了先

证者的那条染色体。

4、观测状态集为：$V=\{0,1\}$。1 代表遗传给先证者的染色体与胚胎遗传草图基因型一致。0 代表不一致。

第一步，构建初始状态概率分布向量，单倍体重组转移矩阵以及观察序列概率矩阵

I. 初始状态概率分布记为 $\pi=\{\pi_i\},(i\in S,\pi_i=0.5)$

在没有参考数据的情况下，设每种隐藏状态出现的可能性相等。

II. 记单倍体重组转移矩阵记为 $A=\{a_{ij}\},(i,j\in S)$,即隐藏状态转移的概率。

$$a_{ij}=\begin{cases}1-Nr/Np & i=j\\ Nr/Np & i\neq j\end{cases}$$

Nr, Np 分别表示期望重组数和单核苷酸多态性位点数。

构建观察序列概率矩阵

记观察序列概率矩阵为 $B=\{b_i(k)\},(i\in S,k\in V)$，这里的概率分布用子代一定是纯合（Must-hom）或一定是杂合(Must-het)的位点来估算。

$$b_i(0)=\frac{\#sites(L>0,\text{Must-hom.})}{\#sites(L>0,\text{Must-het. or Must-hom.})},\ b_i(1)=1-b_i(0)$$

第二步，构建局部概率矩阵，和逆向指针

定义 局部概率 ： $\delta_t(j)=\max\left[\delta_{t-1}(i)\cdot a_{ij}\right]\cdot b_j(K)\quad t\in\{1...N\}$

定义 逆向指针 ： $\Psi_i(j)=\arg\max\delta_{t-1}(i)\cdot a_{ij}\quad t\in\{1...N\}$

第三步，确定最终状态状态，并回溯最优路径

确定最终状态， $q_N^*=\arg\max_{q_N\in S}\delta_N(q_N)$

回溯最优路径，即最可能胚胎单基因型 $q_t^*=\Psi_{t+1}(q^*_{t+1})\ (t=1,2,3,...,N-1)$

第四步，输出结果

实施例 1

一、样品收集及处理:

一对夫妇曾生育过一个重度β地中海贫血的女儿 A(先证者)，因此接受 IVF（体外受精联合胚胎移植技术）-PGD 治疗，希望能再生育一个与先证者 HLA 配型相符的健康宝宝。经过一个 IVF-PGD 治疗周期，在第三天进行卵裂球胚胎单细胞活检取样，活检后的卵裂球继续培养至第五天囊胚期，进行囊胚玻璃化冷冻保存。活检获得的卵裂球单细胞样本进行

全基因组扩增(PED7-1-OWGA)，全基因组扩增采用 Qiagen 公司的 REPLI-g MDA 扩增试剂盒，严格根据试剂盒说明书进行操作，扩增产物用于新一代测序及其他分析。同时也收集这个家系中父亲(PED4-F)、母亲(PED3-M)和先证者女儿(PED-Sister)的外周血，提取基因组 DNA。

在完成 IVF-PGD 治疗后，选择正确的胚胎植入子宫，成功妊娠后，可以在 14 周以后抽取孕妇羊水(PED7-2-OAF)，或收集胚胎脐带血，提取 DNA，进行测序，以验证胚胎细胞的分型结果。

胚胎单细胞全基因组扩增产物 DNA 用 Covaris™ 打断仪打断至 200bp 大小的片段，根据 illumia®公司 HiSeq2000™ 测序仪的上机要求进行建库，构建好的文库经 Agilent®Bioanalyzer 2100 和 Q-PCR 方法进行质控，合格的文库继续进行 illumina® HiSeq2000™ 测序仪测序，测序策略为 Pair End 90 index（即双向 90bp index 测序），测序深度 30X，全基因组测序。其中仪器的参数设置及操作方法都按照 illumina®操作手册（可由 http://www.illumina.com/support/documentation.ilmn 获取）。

父亲、母亲和先证者女儿外周血 DNA，以及孕妇羊水或胚胎脐带血 DNA 用 Covaris™ 打断仪打断至 200bp 大小的片段，根据 illumina®公司 HiSeq2000™ 测序仪的上机要求进行建库，文库进行 All in One 芯片捕获，具体操作见 NimbleGen SeqCap EZ 芯片捕获建库说明书(http://www.nimblegen.com/products/seqcap/ez/index.html 可获取)。All in One 芯片是自主设计的目标区域捕获芯片，其目标区域包括全外显子区域、1M 的 SNP 位点区域和 MHC 基因区。芯片捕获后的文库经 Agilent®Bioanalyzer 2100 和 Q-PCR 方法进行质控，合格的文库继续进行 illumina® HiSeq2000™ 测序仪测序，测序策略为 Pair End 90 index，测序深度为目标区域的 30-50X。其中仪器的参数设置及操作方法都按照 illumina®操作手册（可由 http://www.illumina.com/support/documentation.ilmn 获取）。

二、测序数据分析:

1、先证者和父母基因组 DNA 测序分型:

a) 过滤掉低质量值和接头污染的序列

b) 使用 Burrows-Wheeler Aligner (BWA, http://bio-bwa.sourceforge.net/)将高质量的测序序列比对在人类参考基因组（Hg19, NCBI release GRCh37)。并用 SAMtools 去除掉 PCR 重复扩增的序列。

c) 使用 Genome Analysis Toolkit (GATK, http://www.broadinstitute.org/gatk/)校准碱基质量值并进行单碱基核苷酸多态性和缺失重复的检测。

2、胚胎基因组遗传草图构建。

3、根据先证者基因型推断父母的单倍体型:

根据核心家系推断单体型，具体可参考文献 B. L. Browning and S. R. Browning. Am J Hum Genet 84:210-223.(2009)进行。

4、根据父母单体型推断胚胎细胞的单倍体型：

a) 使用 BWA 将胚胎单细胞测序数据比对到人类参考基因组（Hg19, NCBI release GRCh37)

b) 构建初始状态概率向量，以及胚胎单倍体重组转移矩阵，Nr=30，其它按上所述。

c) 统计胚胎细胞每个位点的测序信息，并构建观察序列概率矩阵（按上所述）。

d) 构建局部概率矩阵，和逆向指针（按上所述）。

e) 确定最终状态，并回溯最优路径

f) 输出

g) 以胚胎出生后的羊水基因分型结果为参照，对胚胎细胞的基因分型准确性统计如下：

种类		常染色体			性染色体		
		准确数	位点数	准确率	准确数	位点数	准确率
重构前	纯合	175,194	233,794	74.94	2,345	3,318	70.68
	杂合	81,129	81,576	99.45	1,476	1,480	99.72
重构后	纯合	172,968	173,631	99.62	2,329	2,330	99.96
	草图杂合	80,877	81,042	99.80	1,472	1,478	99.59
	更正杂合	58,194	60,194	95.88	967	990	97.68

注："草图杂合"表示重构前后均为杂合的位点；"更正杂合"表示重构前为纯合，重构后为杂合的位点。

5、β-地贫和 HLA 型别判断

a) 通过父母单体型，成功还原了胚胎细胞在 rs7480526 上的基因型，预测其为β-地贫阴性，并发现与先证者在该位置上的等位基因不一样，从而确定胚胎为β-地贫阴性。

b) 通过本方法可以得出，先证者和胚胎在 MHC 区的单体型一样，从而判断她们的 HLA 型别是匹配的。

根据上述胚胎的β-地贫基因型和 HLA 分型结果，从而选择正确的胚胎，即β-地贫阴性和 HLA 型别与先证者匹配的胚胎，进行冷冻复苏和植入子宫，从而防止患病儿的妊娠。

实施例 2

一、样品收集及处理：

一对夫妇均为β-地贫携带者，接受 IVF-PGD 治疗。经过一个 IVF-PGD 治疗周期，在第三天获得了 8 个卵裂球胚胎，均进行卵裂球胚胎单细胞活检取样，活检后的卵裂球继续培养至第五天囊胚期，进行囊胚玻璃化冷冻保存。活检获得的 8 个卵裂球单细胞样本进行全基因组扩增，全基因组扩增采用 Qiagen 公司的 REPLI-g MDA 扩增试剂盒，严格根据试剂盒说明书进行操作，扩增产物用于新一代测序及其他分析。

同时也收集这个家系中父亲、母亲的外周血，提取基因组 DNA。

在完成 IVF-PGD 治疗，选择正确的胚胎植入子宫，成功妊娠后，可以在 14 周以后抽取孕妇羊水，或收集胚胎脐带血，提取 DNA，进行测序，以验证胚胎细胞的分型结果。

8 个胚胎单细胞全基因组扩增产物 DNA，父亲、母亲外周血 DNA，以及孕妇羊水或胚胎脐带血 DNA 用 Covaris™ 打断仪打断至 200bp 大小的片段，根据 illumina®公司 HiSeq2000™ 测序仪的上机要求进行建库，文库进行 All in One 芯片捕获，具体操作见 NimbleGen SeqCap EZ 芯片捕获建库说明书 (http://www.nimblegen.com/products/seqcap/ez/index.html 可获取)。All in One 芯片是自主设计的目标区域捕获芯片，其目标区域包括全外显子区域、1M 的 SNP 位点区域和 MHC 基因区。芯片捕获后的文库经 Agilent®Bioanalyzer 2100 和 Q-PCR 方法进行质控，合格的文库继续进行 illumina® HiSeq2000™ 测序仪测序，测序策略为 Pair End 90 index（即双向 90bp index 测序），测序深度为目标区域的 30-50X。其中仪器的参数设置及操作方法都按照 illumina® 操作手册（可由 http://www.illumina.com/support/documentation.ilmn 获取）。

二、测序数据分析:

1) 父母基因组 DNA，羊水 DNA（验证样本）测序数据分析和分型:

a) 过滤掉低质量值和接头污染的序列

b) 使用 Burrows-Wheeler Aligner (BWA, http://bio-bwa.sourceforge.net/)将高质量的测序序列比对在人类参考基因组（Hg19, NCBI release GRCh37)。并用 SAMtools 去除掉 PCR 重复扩增的序列。

c) 使用 Genome Analysis Toolkit (GATK, http://www.broadinstitute.org/gatk/)校准碱基质量值并进行单碱基核苷酸多态性和缺失重复的检测。

2) 根据 8 个胚胎单细胞测序数据分析，分别给出每个胚胎细胞基因组初始化的基因型信息，构建每个胚胎的细胞基因组遗传草图。

3）确定 8 个胚胎单细胞任一个待测胚胎，根据待测胚胎细胞基因组遗传草图和父母的基因型，对父母的单倍体型进行初始化。

4）借助隐马尔科夫模型，分别根据其余 7 个胚胎的基因型重构并确定父母的单倍体型。

5）借助隐马尔科夫模型，根据父母单倍体型，来重构待测胚胎的单体型。

6）待测胚胎遗传变异信息的确定。

以 22 号染色体为例，参照羊水基因分型结果，对待测胚胎的基因分型准确性统计如下:

类型(母 x 父)	总位点数	SNP Call 率	准确率(杂合)	准确率(纯合)
纯合 x 杂合	2183	97.16%	96.68%	99.63%
杂合 x 纯合	2156	93.18%	98.95%	100%
杂合 x 杂合	1154	82.32%	99.31%	97.59%

7) β-地贫和 HLA 型别判断

a) 通过父母单体型，成功还原了待测胚胎细胞在 rs7480526 上的单体型，预测其为 β-地贫阴性。

b) 通过本方法，结合先症者的遗传检测结果，可判定先证者和胚胎在 MHC 区的单体型是否一致，从而判断她们的 HLA 型别是否匹配。

根据上述胚胎的β-地贫基因型和 HLA 分型结果，从而选择正确的胚胎，即β-地贫阴性和 HLA 型别与先症者匹配的胚胎，进行冷冻复苏和植入子宫，从而防止患病儿的妊娠。

工业实用性

本发明的确定胚胎基因组中预定区域碱基信息的方法、用于确定胚胎基因组中预定区域碱基信息的系统以及计算机可读介质，能够有效地应用于对胚胎基因组中预定区域的核酸序列进行分析。

尽管本发明的具体实施方式已经得到详细的描述，本领域技术人员将会理解。根据已经公开的所有教导，可以对那些细节进行各种修改和替换，这些改变均在本发明的保护范围之内。本发明的全部范围由所附权利要求及其任何等同物给出。

在本说明书的描述中，参考术语“一个实施例”、“一些实施例”、“示意性实施例”、“示例”、“具体示例”、或“一些示例”等的描述意指结合该实施例或示例描述的具体特征、结构、材料或者特点包含于本发明的至少一个实施例或示例中。在本说明书中，对上述术语的示意性表述不一定指的是相同的实施例或示例。而且，描述的具体特征、结构、材料或者特点可以在任何的一个或多个实施例或示例中以合适的方式结合。

权利要求书

1、一种确定胚胎基因组中预定区域碱基信息的方法，其特征在于，包括下列步骤：

获得胚胎细胞基因组 DNA 样本的测序结果，以及胚胎遗传相关个体基因组样本的测序结果；

基于胚胎细胞基因组 DNA 样本的测序结果，构建所述胚胎的遗传草图，以便确定胚胎初始基因型；

基于所述胚胎遗传相关个体基因组样本的测序结果，确定胚胎父母的单倍体型；以及

根据隐马尔可夫模型，以所述胚胎初始基因型作为观察序列，基于所述胚胎父母的单倍体型，确定胚胎基因组中预定区域的碱基信息。

2、根据权利要求 1 所述的方法，其特征在于，所述胚胎基因组 DNA 样本是从微量胚胎细胞获得的，其中，所述微量胚胎细胞来自于卵裂球、囊胚或受精卵。

3、根据权利要求 2 所述的方法，其特征在于，所述胚胎基因组 DNA 样本是从胚胎单细胞获得的。

4、根据权利要求 1 所述的方法，其特征在于，所述测序结果是利用选自 Illumina-Solexa/Hiseq、ABI-SOLiD、Life Technologies-Ion Torrent/Proton、Roche-454 和单分子测序装置的至少一种获得的。

5、根据权利要求 1 所述的方法，其特征在于，通过将胚胎细胞基因组 DNA 样本的测序结果与参考序列进行比对，构建所述胚胎的遗传草图。

6、根据权利要求 1 所述的方法，其特征在于，通过将所述胚胎遗传相关个体基因组样本的测序结果与参考序列进行比对，确定所述胚胎遗传相关个体的基因型；以及

基于所述胚胎遗传相关个体的基因型，确定所述胚胎父母的单倍体型。

7、根据权利要求 5 或 6 所述的方法，其特征在于，所述参考序列为人类参考基因组。

8、根据权利要求 1 所述的方法，其特征在于，所述胚胎遗传相关个体包括所述胚胎的父母以及先证者。

9、根据权利要求 8 所述的方法，其特征在于，所述先证者为所述胚胎父母的其他子女。

10、根据权利要求 9 所述的方法，其特征在于，根据隐马尔可夫模型，确定胚胎基因组中预定区域的碱基信息进一步包括：

构建初始状态概率分布向量、隐藏状态转移的概率矩阵以及观察序列概率矩阵；

利用惠特比算法确定最终状态并回溯最优路径，以便确定胚胎基因组中预定区域的碱基信息。

11、根据权利要求 10 所述的方法，其特征在于，所述隐马尔可夫模型采用下列参数：

初始状态概率分布为 $\pi = \{\pi_i\}, (i \in S, \pi_i = 0.5)$，

隐藏状态转移的概率矩阵为 $A=\{a_{ij}\},(i,j\in S)$，其中，$a_{ij}=\begin{cases}1-Nr/Np & i=j\\ Nr/Np & i\neq j\end{cases}$

Nr、Np 分别表示期望重组数和单核苷酸多态性位点数，Nr 为自然数，取值范围 20-40，观察序列概率矩阵为 $B=\{b_i(k)\},(i\in S,k\in V)$，其中，$b_i(0)=\frac{\#sites(L>0,\text{Must-hom.})}{\#sites(L>0,\text{Must-het. or Must-hom.})}$，$b_i(1)=1-b_i(0)$，

#sites (L>0,Must-hom.)为子代一定为纯合的位点的数目，#sites (L>0,Must-hom. or Must-het.)为子代一定为纯合的位点的数目以及子代一定为杂合的位点的数目的总和；

局部概率为 $\delta_t(j)=\max\left[\delta_{t-1}(i)\cdot a_{ij}\right]\cdot b_j(K)\quad t\in\{1...N\}$，

逆向指针为 $\Psi_t(j)=\arg\max\delta_{t-1}(i)\cdot a_{ij}\quad t\in\{1...N\}$，

递归得最终状态为 $q_N^*=\arg\max_{q_N\in S}\delta_N(q_N)$，

回溯最优路径，确定最可能的胚胎预定区域的碱基信息为 $q_t^*=\Psi_{t+1}(q^*_{t+1})$ $(t=1,2,3,...,N-1)$。

12、根据权利要求 1 所述的方法，其特征在于，进一步包括：所述预定区域为已知存在遗传多态性的位点。

13、根据权利要求 12 所述的方法，其特征在于，所述遗传多态性为选自单核苷酸多态性和 STR 的至少一种。

14、一种用于确定胚胎基因组中预定区域碱基信息的系统，其特征在于，包括：

文库构建装置，所述文库构建装置适于针对胚胎基因组 DNA 样本以及胚胎遗传相关个体的基因组 DNA 样本，分别构建测序文库；

测序装置，所述测序装置与所述文库构建装置相连，并且适于对所述测序文库进行测序，以便获得胚胎以及胚胎遗传相关个体的测序结果；以及

分析装置，所述分析装置与所述测序装置相连，并且适于：

基于胚胎基因组 DNA 样本的测序结果，构建所述胚胎的遗传草图，以便确定胚胎初始基因型；

基于所述胚胎遗传相关个体基因组样本的测序结果，确定胚胎父母的单倍体型；以及

根据隐马尔可夫模型，以所述胚胎初始基因型作为观察序列，基于所述胚胎父母的单倍体型，确定胚胎基因组中预定区域的碱基信息。

15、根据权利要求 14 所述的系统，其特征在于，进一步包括 DNA 样本分离装置，所述 DNA 样本分离装置适于从胚胎进行活检取样，获得微量胚胎细胞，从所述微量胚胎细胞

提取胚胎基因组DNA样本，其中，所述微量胚胎细胞来自于卵裂球、囊胚或受精卵。

16、根据权利要求14所述的系统，其特征在于，所述测序装置为选自Illumina-Solexa/Hiseq、ABI-SOLiD、Life Technologies-Ion Torrent/Proton、Roche-454和单分子测序装置的至少一种。

17、根据权利要求14所述的系统，其特征在于，进一步包括比对装置，所述比对装置与所述测序装置相连，用于将所述测序结果与参考序列进行比对，以便：

通过将胚胎细胞基因组DNA样本的测序结果与参考序列进行比对，构建所述胚胎的遗传草图；

通过将所述胚胎遗传相关个体基因组样本的测序结果与参考序列进行比对，确定所述胚胎遗传相关个体的基因型；以及

基于所述胚胎遗传相关个体的基因型，确定所述胚胎父母的单倍体型。

18、根据权利要求14所述的系统，其特征在于，根据隐马尔可夫模型，确定胚胎基因组中预定区域的碱基信息进一步包括：

构建初始状态概率分布向量、隐藏状态转移的概率矩阵以及观察序列概率矩阵；

利用惠特比算法确定最终状态并回溯最优路径，以便确定胚胎基因组中预定区域的碱基信息。

19、根据权利要求18所述的系统，其特征在于，所述隐马尔可夫模型采用下列参数：

初始状态概率分布为 $\pi=\{\pi_i\},(i\in S,\pi_i=0.5)$，

隐藏状态转移的概率矩阵为 $A=\{a_{ij}\},(i,j\in S)$，其中，$a_{ij}=\begin{cases}1-Nr/Np & i=j\\ Nr/Np & i\neq j\end{cases}$

Nr、Np分别表示期望重组数和单核苷酸多态性位点数，Nr为自然数，取值范围20-40，

观察序列概率矩阵为 $B=\{b_i(k)\},(i\in S,k\in V)$，其中，

$b_i(0)=\dfrac{\#sites(L>0,\text{Must-hom.})}{\#sites(L>0,\text{Must-het. or Must-hom.})}$，$b_i(1)=1-b_i(0)$，

#sites (L>0,Must-hom.)为子代一定为纯合的位点的数目，#sites (L>0,Must-hom. or Must-het.)为子代一定为纯合的位点的数目以及子代一定为杂合的位点的数目的总和；

局部概率为 $\delta_t(j)=\max\left[\delta_{t-1}(i)\cdot a_{ij}\right]\cdot b_j(K)\quad t\in\{1...N\}$，

逆向指针为 $\Psi_t(j)=\arg\max\delta_{t-1}(i)\cdot a_{ij}\quad t\in\{1...N\}$，

递归得最终状态为 $q_N^*=\arg\max_{q_N\in S}\delta_N(q_N)$，

回溯最优路径，确定最可能的胚胎预定区域的碱基信息为 $q_t^*=\Psi_{t+1}(q^*_{t+1})$

($t=1,2,3,...,N-1$)。

20、一种计算机可读介质，其特征在于，所述计算机可读介质上存储有指令，所述指令适于被处理器执行以便:

基于胚胎细胞基因组 DNA 样本的测序结果，构建胚胎的遗传草图，以便确定胚胎初始基因型;

基于所述胚胎遗传相关个体基因组样本的测序结果，确定胚胎父母的单倍体型；以及

根据隐马尔可夫模型，以所述胚胎初始基因型作为观察序列，基于所述胚胎父母的单倍体型，确定胚胎基因组中预定区域的碱基信息。

21、根据权利要求 20 所述的计算机介质，其特征在于，通过将胚胎细胞基因组 DNA 样本的测序结果与参考序列进行比对，构建所述胚胎的遗传草图。

22、根据权利要求 20 所述的计算机介质，其特征在于，通过将所述胚胎遗传相关个体基因组样本的测序结果与参考序列进行比对，确定所述胚胎遗传相关个体的基因型；以及

基于所述胚胎遗传相关个体的基因型，确定所述胚胎父母的单倍体型。

23、根据权利要求 20 所述的计算机介质，其特征在于，根据隐马尔可夫模型，确定胚胎基因组中预定区域的碱基信息进一步包括:

构建初始状态概率分布向量、隐藏状态转移的概率矩阵以及观察序列概率矩阵;

利用惠特比算法确定最终状态并回溯最优路径，以便确定胚胎基因组中预定区域的碱基信息。

24、根据权利要求 23 所述的方法，其特征在于，所述隐马尔可夫模型采用下列参数:

初始状态概率分布为 $\pi=\{\pi_i\},(i\in S,\pi_i=0.5)$,

隐藏状态转移的概率矩阵为 $A=\{a_{ij}\},(i,j\in S)$，其中，$a_{ij}=\begin{cases}1-Nr/Np & i=j\\ Nr/Np & i\neq j\end{cases}$

Nr、Np 分别表示期望重组数和单核苷酸多态性位点数，Nr 为自然数，取值范围 20-40，

观察序列概率矩阵为 $B=\{b_i(k)\},(i\in S,k\in V)$，其中，

$b_i(0)=\frac{\#sites(L>0,\text{Must-hom.})}{\#sites(L>0,\text{Must-het. or Must-hom.})}$, $b_i(1)=1-b_i(0)$,

#sites (L>0,Must-hom.)为子代一定为纯合的位点的数目，#sites (L>0,Must-hom. or Must-het.)为子代一定为纯合的位点的数目以及子代一定为杂合的位点的数目的总和;

局部概率为 $\delta_t(j)=\max\left[\delta_{t-1}(i)\cdot a_{ij}\right]\cdot b_j(K)\quad t\in\{1...N\}$,

逆向指针为 $\Psi_i(j)=\arg\max\delta_{t-1}(i)\cdot a_{ij}\quad t\in\{1...N\}$,

递归得最终状态为 $q_N^*=\arg\max_{q_N\in S}\delta_N(q_N)$,

回溯最优路径，确定最可能的胚胎预定区域的碱基信息为 $q_t^* = \Psi_{t+1}(q^*_{t+1})$ ($t = 1,2,3,...,N-1$)。

25、根据权利要求 20 所述的计算机可读介质，其特征在于，所述预定区域为已知存在遗传多态性的位点。

26、根据权利要求 25 所述的计算机可读介质，其特征在于，所述遗传多态性为选自单核苷酸多态性和 STR 的至少一种。

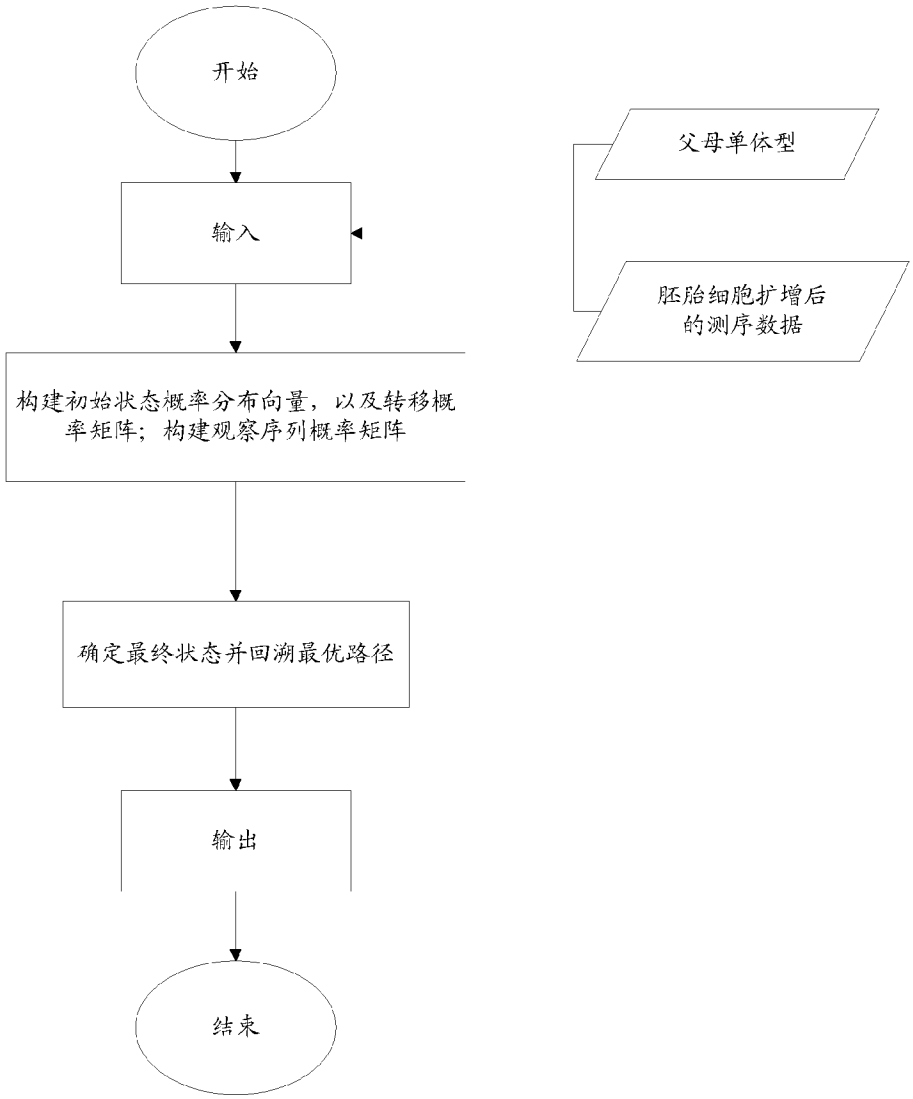

开始
输入
父母单体型
胚胎细胞扩增后的测序数据
构建初始状态概率分布向量，以及转移概率矩阵；构建观察序列概率矩阵
确定最终状态并回溯最优路径
输出
结束

图 1

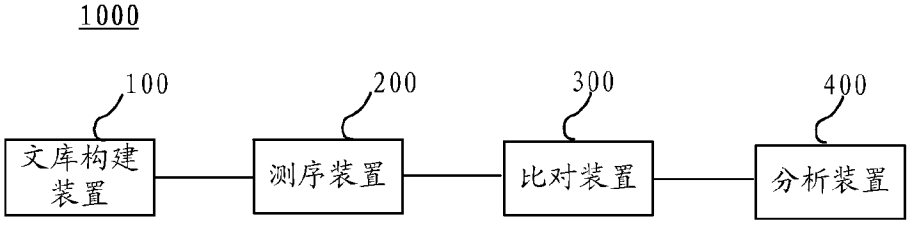

1000
100
200
300
400
文库构建装置
测序装置
比对装置
分析装置

图 2

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2013/073375

A. CLASSIFICATION OF SUBJECT MATTER

C12Q 1/68 (2006.01) i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

IPC: C12Q

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

CNPAT, WPI, EPODOC, CNKI, PubMed, GOOGLE, NCBI: hidden markov model, HMM, Viterbi algorithm, library, sequencing, hidden, embryo, creat

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	YU, Weidong et al. Construction of cDNA Library from Rabbit Single Preimplantation Cloned Embryos. Chinese Journal of Biochemistry and Molecular Biology, April 2004, vol. 20, no. 2, pages 162-165, page 163, right columns 1.2.3 and 1.2.4	1-19
Y	CN 102317473 A (PACIFIC BIOSCIENCES CALIFORNIA) 11 January 2012 (11.01.2012) see claims 1 and 2, description, paragraphs [0161]-[0166]	1-19
A	CN 102127818 A (ZHANG, Kang) 20 July 2011 (20.07.2011) see the whole document	1-19

☐ Further documents are listed in the continuation of Box C. ☒ See patent family annex.

* Special categories of cited documents:

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier application or patent but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
10 December 2013 (10.12.2013)	16 January 2014 (16.01.2014)

Name and mailing address of the ISA	Authorized officer
State Intellectual Property Office of the P. R. China No. 6, Xitucheng Road, Jimenqiao Haidian District, Beijing 100088, China Facsimile No. (86-10) 62019451	ZHANG, Bin Telephone No. (86-10) 61648321

Form PCT/ISA/210 (second sheet) (July 2009)

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2013/073375

Box No. II Observations where certain claims were found unsearchable (Continuation of item 2 of first sheet)

This international search report has not been established in respect of certain claims under Article 17(2)(a) for the following reasons:

1. ☒ Claims Nos.:20-26
because they relate to subject matter not required to be searched by this Authority, namely:
The subject matter of claims 20-26 relates to mere presentations of information and therefore does not warrant an international search according to the criteria set out in PCT Rule 39. 1(iv).

2. ☐ Claims Nos.:
because they relate to parts of the international application that do not comply with the prescribed requirements to such an extent that no meaningful international search can be carried out, specifically:

3. ☐ Claims Nos.:
because they are dependent claims and are not drafted in accordance with the second and third sentences of Rule 6.4(a).

Box No. III Observations where unity of invention is lacking (Continuation of item 3 of first sheet)

This International Searching Authority found multiple inventions in this international application, as follows:

1. ☐ As all required additional search fees were timely paid by the applicant, this international search report covers all searchable claims.

2. ☐ As all searchable claims could be searched without effort justifying additional fees, this Authority did not invite payment of additional fees.

3. ☐ As only some of the required additional search fees were timely paid by the applicant, this international search report covers only those claims for which fees were paid, specifically claims Nos.:

4. ☐ No required additional search fees were timely paid by the applicant. Consequently, this international search report is restricted to the invention first mentioned in the claims; it is covered by claims Nos.:

Remark on protest

☐ The additional search fees were accompanied by the applicant's protest and, where applicable, the payment of a protest fee.

☐ The additional search fees were accompanied by the applicant's protest but the applicable protest fee was not paid within the time limit specified in the invitation.

☐ No protest accompanied the payment of additional search fees.

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2013/073375

Patent Documents referred in the Report	Publication Date	Patent Family	Publication Date
CN 102317473 A	11.01.2012	WO 2010068289 A2	17.06.2010
		US 2011183320 A1	28.07.2011
		US 2010221716 A1	02.09.2010
		CA 2746632 A1	17.06.2010
		WO 2010068289 A3	14.10.2010
		AU 2009325069 A1	17.06.2010
		EP 2370598 A2	05.10.2011
		US 2013029853 A1	31.01.2013
CN 102127818 A	20.07.2011	None	

Form PCT/ISA /210 (patent family annex) (July 2009)

国际检索报告

国际申请号
PCT/CN2013/073375

A. 主题的分类

C12Q 1/68 (2006.01)i

按照国际专利分类(IPC)或者同时按照国家分类和 IPC 两种分类

B. 检索领域

检索的最低限度文献(标明分类系统和分类号)

IPC : C12Q

包含在检索领域中的除最低限度文献以外的检索文献

在国际检索时查阅的电子数据库(数据库的名称，和使用的检索词（如使用））

CNPAT, WPI, EPODOC, CNKI, PubMed, GOOGLE, NCBI：隐马尔可夫模型，惠特比算法，文库，测序，隐马，胚胎，构建，测序，library，sequencing，hidden，HMM.

C. 相关文件

类　型*	引用文件，必要时，指明相关段落	相关的权利要求
Y	郁卫东 等，兔单个植入前克隆胚胎 cDNA 文库的构建，中国生物化学与分子生物学报，04 月 2004，第 20 卷，第 2 期，第 162-165 页 参见第 163 页右栏 1.2.3 和 1.2.4	1-19
Y	CN102317473 A（加利福尼亚太平洋生物科学股份有限公司）11.01 月 2012（11.01.2012） 参见权利要求 1-2，说明书第[0161]-[0166]段	1-19
A	CN102127818 A（张康）20.07 月 2011（20.07.2011） 参见全文	1-19

☐ 其余文件在 C 栏的续页中列出。　　☒ 见同族专利附件。

* 引用文件的具体类型：
"A" 认为不特别相关的表示了现有技术一般状态的文件
"E" 在国际申请日的当天或之后公布的在先申请或专利
"L" 可能对优先权要求构成怀疑的文件，或为确定另一篇引用文件的公布日而引用的或者因其他特殊理由而引用的文件（如具体说明的）
"O" 涉及口头公开、使用、展览或其他方式公开的文件
"P" 公布日先于国际申请日但迟于所要求的优先权日的文件
"T" 在申请日或优先权日之后公布，与申请不相抵触，但为了理解发明之理论或原理的在后文件
"X" 特别相关的文件，单独考虑该文件，认定要求保护的发明不是新颖的或不具有创造性
"Y" 特别相关的文件，当该文件与另一篇或者多篇该类文件结合并且这种结合对于本领域技术人员为显而易见时，要求保护的发明不具有创造性
"&" 同族专利的文件

国际检索实际完成的日期	国际检索报告邮寄日期
10.12 月 2013（10.12.2013）	16.1 月 2014 (16.01.2014)
ISA/CN 的名称和邮寄地址： 中华人民共和国国家知识产权局 中国北京市海淀区蓟门桥西上城路 6 号 100088 传真号：　(86-10)62019451	受权官员 张彬 电话号码：(86-10) **61648321**

PCT/ISA/210 表(第 2 页) (2009 年 7 月)

国际检索报告

国际申请号
PCT/CN2013/073375

第II栏　某些权利要求被认为是不能检索的意见(续第1页第2项)

根据条约第17条(2)(a)，对某些权利要求未做国际检索报告的理由如下：

1. ☒ 权利要求：20-26

因为它们涉及不要求本单位进行检索的主题，即：

权利要求20-26涉及单纯的信息提供（PCT细则39.1(v)）。

2. ☐ 权利要求：

因为它们涉及国际申请中不符合规定的要求的部分，以致不能进行任何有意义的国际检索，

具体地说：

3. ☐ 权利要求：

因为它们是从属权利要求，并且没有按照细则6.4(a)第2句和第3句的要求撰写。

第III栏　缺乏发明单一性的意见(续第1页第3项)

本国际检索单位在该国际申请中发现多项发明，即：

1. ☐ 由于申请人按时缴纳了被要求缴纳的全部附加检索费，本国际检索报告涉及全部可作检索的权利要求。

2. ☐ 由于无需付出有理由要求附加费的劳动即能对全部可检索的权利要求进行检索，本单位未通知缴纳任何附加费。

3. ☐ 由于申请人仅按时缴纳了部分被要求缴纳的附加检索费，本国际检索报告仅涉及已缴费的那些权利要求。具体地说，是权利要求：

4. ☐ 申请人未按时缴纳被要求缴纳的附加检索费。因此，本国际检索报告仅涉及权利要求书中首先提及的发明；包含该发明的权利要求是：

关于异议的说明： ☐ 申请人缴纳了附加检索费，同时提交了异议书，适用时，缴纳了异议费。

☐ 申请人缴纳了附加检索费，同时提交了异议书，但未在通知书规定的时间期限内缴纳异议费。

☐ 缴纳附加检索费时未提交异议书。

国际检索报告
关于同族专利的信息

国际申请号
PCT/CN2013/073375

检索报告中引用的专利文件	公布日期	同族专利	公布日期
CN 102317473 A	11.01.2012	WO 2010068289 A2	17.06.2010
		US 2011183320 A1	28.07.2011
		US 2010221716 A1	02.09.2010
		CA 2746632 A1	17.06.2010
		WO 2010068289 A3	14.10.2010
		AU 2009325069 A1	17.06.2010
		EP 2370598 A2	05.10.2011
		US 2013029853 A1	31.01.2013
CN 102127818 A	20.07.2011	无	

(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
19 December 2013 (19.12.2013)

WIPO | PCT

(10) International Publication Number
WO 2013/188846 A1

(51) International Patent Classification:
C12Q 1/68 (2006.01) *C40B 30/04* (2006.01)
G01N 33/53 (2006.01)

(21) International Application Number:
PCT/US2013/046020

(22) International Filing Date:
14 June 2013 (14.06.2013)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
61/660,427 15 June 2012 (15.06.2012) US

(72) Inventor; and
(71) Applicant : **STYLLI, Harry** [GB/US]; 2452 Pasco Dorado, La Jolla, CA 92037 (US).

(74) Agents: **HALEY, James, F.** et al.; Ropes & Gray LLP, 1211 Avenue of the Americas, New York, NY 10036 (US).

(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IS, JP, KE, KG, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(84) Designated States *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

Published:

— *with international search report (Art. 21(3))*

— *before the expiration of the time limit for amending the claims and to be republished in the event of receipt of amendments (Rule 48.2(h))*

(54) Title: METHODS OF DETECTING DISEASES OR CONDITIONS

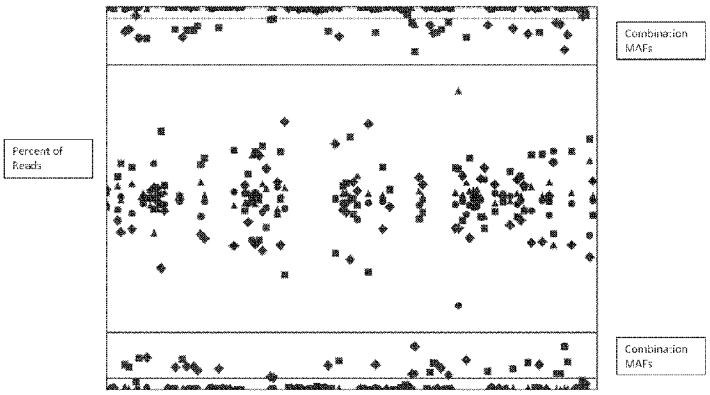


Figure 2

(57) Abstract: This invention provides methods of using a sample with multiple analytical components in the diagnosis, prognosis, or monitoring of diseases or conditions. The invention also provides methods of identifying markers of diseases or conditions.

WO 2013/188846 A1

METHODS OF DETECTING DISEASES OR CONDITIONS

Priority Information

[0001] This application claims priority from U.S. Provisional Application 61/660,427, filed June 15, 2012. The content and disclosure of that application is incorporated by reference herein in its entirety.

Field of the Invention

[0002] This invention relates generally to methods of using combinations of two or more different components selected from cell-free bodily fluids, phagocytic cells, circulating vesicles, and circulating diseased cells in the diagnosis, prognosis, or monitoring of a disease or condition. The invention also relates to methods of using the combinations to identify markers of diseases or conditions.

Background of the Invention

[0003] Leukocytes begin as pluripotent hematopoietic stem cells in the bone marrow and develop along either the myeloid lineage (monocytes, macrophages, neutrophils, eosinophils, and basophils) or the lymphoid lineage (T and B lymphocytes and natural killer cells). The major function of the myeloid lineage cells (e.g., neutrophils and macrophages) is the phagocytosis of infectious organisms, live unwanted damaged cells, senescent and dead cells (apoptotic and necrotic), as well as the clearing of cellular debris. Phagocytes from healthy animals do not replicate and are diploid, i.e., have a DNA content of 2n. On average, each cell contains <10 ng

DNA, <20 ng RNA, and <300 ng of protein. Non-phagocytic cells are also diploid and are not involved in the internalization of dead cells or infectious organisms and also have a DNA content of 2n.

[0004] The lifetime of various white blood cell subpopulations varies from a few days (e.g., neutrophils) to several months (e.g., macrophages). Like other cell types, leukocytes age and eventually die. During their aging process, human blood- and tissue-derived phagocytes (e.g., neutrophils) exhibit all the classic markers of programmed cell death (i.e., apoptosis), including caspase activation, pyknotic nuclei, and chromatin fragmentation. These cells also display a number of "eat-me" flags (e.g., phosphatidylserine, sugars) on the extracellular surfaces of their plasma membranes. Consequently, dying and dead cells and subcellular fragments thereof are cleared from tissues and blood by other phagocytic cells.

[0005] Early diagnosis of a disease often increases the likelihood of successful treatment or cure of such disease. Current diagnostic methods, however, focus on either using whole blood or separating the blood into different components, of which a single component is chosen for testing. Although this approach may enrich the signal being detected, it also results in the loss of potentially important information. Personalized diagnostic methods are needed that enable the diagnosis, especially the early diagnosis, of the presence of a disease or a condition in individuals who are not known to have the disease or who have recurrent disease.

[0006] One object of the present invention is to provide diagnostic methods that can facilitate the detection of a disease or condition-specific markers, e.g., nucleic acids, proteins, carbohydrates, and/or lipids and the like by using combinations of two or more different components selected from cell-free bodily fluids, phagocytic cells, circulating vesicles, and circulating diseased cells. Another object of this invention is to provide methods of identifying a disease or condition-specific markers and further use such markers alone or together with any known markers to diagnose diseases or conditions.

Summary of the Invention

[0007] Some embodiments of the invention are:

1. A method for diagnosing or aiding in the diagnosis of a disease or condition in a subject comprising:

a) determining a first profile of one or more markers of the disease or condition from a sample comprising two or more different components selected from the group consisting of: a cell-free bodily fluid isolated from the subject, a population of phagocytic cells isolated from the subject, a population of phagocytic cells having a DNA content more than 2n (>2n phagocytic cells) isolated from the subject, a population of circulating vesicles isolated from the subject, and a population of circulating diseased cells isolated from the subject;

b) determining a second profile of at least one of the one or more markers from a control comprising a component selected from the group consisting of: a population of phagocytic cells having a DNA content of 2n (=2n phagocytic cells) isolated from the subject, a population of non-phagocytic cells isolated from the subject, and a population of control cells isolated from the subject, wherein the control cells are substantially free of cells affected by the disease or condition; and

c) identifying a difference between the first and second profiles, wherein the difference is indicative of the presence of said disease or condition in the subject.

2. A method for assessing the risk of developing a disease or condition in a subject comprising:

a) determining a first profile of one or more markers of the disease or condition from a sample comprising two or more different components selected from the group consisting of: a cell-free bodily fluid isolated from the subject, a population of phagocytic cells isolated from the subject, a population of phagocytic cells having a DNA content more than 2n (>2n phagocytic cells) isolated from the subject, a population of circulating vesicles isolated from the subject, and a population of circulating diseased cells isolated from the subject;

b) determining a second profile of at least one of the one or more markers from a control comprising a component selected from the group consisting of: a population of phagocytic cells having a DNA content of 2n (=2n phagocytic cells) isolated from the subject, a population of non-phagocytic cells isolated from the subject, and a population of control cells isolated from the subject, wherein the control cells are substantially free of cells affected by the disease or condition; and

c) identifying a difference between the first and second profiles, wherein the difference is indicative of the risk of developing said disease or condition in the subject.

3. A method for prognosing or aiding in the prognosis of a disease or condition in a subject comprising:

a) determining a first profile of one or more markers of the disease or condition from a sample comprising two or more different components selected from the group consisting of: a cell-free bodily fluid isolated from the subject, a population of phagocytic cells isolated from the subject, a population of phagocytic cells having a DNA content more than 2n (>2n phagocytic cells) isolated from the subject, a population of circulating vesicles isolated from the subject, and a population of circulating diseased cells isolated from the subject;

b) determining a second profile of at least one of the one or more markers from a control comprising a component selected from the group consisting of: a population of phagocytic cells having a DNA content of 2n (=2n phagocytic cells) isolated from the subject, a population of non-phagocytic cells isolated from the subject, and a population of control cells isolated from the subject, wherein the control cells are substantially free of cells affected by the disease or condition; and

c) identifying a difference between the first and second profiles, wherein the difference is indicative of the prognosis of said disease or condition in the subject.

4. A method for assessing the efficacy of a treatment for a disease or condition in a subject comprising:

a) determining a first profile of one or more markers of the disease or condition from a first sample comprising two or more different components selected from the group consisting of: a cell-free bodily fluid isolated from the subject before the treatment, a population of phagocytic cells isolated from the subject before the treatment, a population of phagocytic cells having a DNA content more than 2n (>2n phagocytic cells) isolated from the subject before the treatment, a population of circulating vesicles isolated from the subject before the treatment, and a population of circulating diseased cells isolated from the subject before the treatment;

determining a second profile of at least one of the one or more markers from a first control comprising a component selected from the group consisting of: a population of phagocytic cells having a DNA content of 2n (=2n phagocytic cells) isolated from the subject before the treatment, a population of non-phagocytic cells isolated from the subject before the treatment, and a population of control cells isolated from the subject, wherein the control cells are substantially free of cells affected by the disease or condition, before the treatment;

identifying a difference between the first and second profiles;

b) determining a third profile of one or more markers of the disease or condition from a second sample comprising two or more different components selected from the group consisting of: a cell-free bodily fluid isolated from the subject after the treatment, a population of phagocytic cells isolated from the subject after the treatment, a population of phagocytic cells having a DNA content more than 2n (>2n phagocytic cells) isolated from the subject after the treatment, a population of circulating vesicles isolated from the subject after the treatment, and a population of circulating diseased cells isolated from the subject after the treatment;

determining a fourth profile of at least one of the one or more markers from a second control comprising a component selected from the group consisting of: a population of phagocytic cells having a DNA content of 2n (=2n phagocytic cells) isolated from the subject after the treatment, a population of non-phagocytic cells isolated from the subject after the treatment, and a population of control cells isolated

from the subject, wherein the control cells are substantially free of cells affected by the disease or condition, after the treatment;

identifying a difference between the third and fourth profiles; and

c) identifying a difference between the difference identified in a) and the difference identified in b) wherein the identified difference in c) is indicative of the efficacy of the treatment for said disease or condition in the subject.

5. A method for monitoring the progression or regression of a disease or condition in a subject comprising:

a) determining a first profile of one or more markers of the disease or condition from a first sample comprising two or more different components selected from the group consisting of: a cell-free bodily fluid isolated from the subject at a first time point, a population of phagocytic cells isolated from the subject at a first time point, a population of phagocytic cells having a DNA content more than 2n (>2n phagocytic cells) isolated from the subject at a first time point, a population of circulating vesicles isolated from the subject at a first time point, and a population of circulating diseased cells isolated from the subject at a first time point;

determining a second profile of at least one of the one or more markers from a first control comprising a component selected from the group consisting of: a population of phagocytic cells having a DNA content of 2n (=2n phagocytic cells) isolated from the subject at a first time point, a population of non-phagocytic cells isolated from the subject at a first time point, and a population of control cells isolated from the subject, wherein the control cells are substantially free of cells affected by the disease or condition, at a first time point;

identifying a difference between the first and second profiles;

b) determining a third profile of one or more markers of the disease or condition from a second sample comprising two or more different components selected from the group consisting of: a cell-free bodily fluid isolated from the subject at a second time point, a population of phagocytic cells isolated from the subject at a

second time point, a population of phagocytic cells having a DNA content more than 2n (>2n phagocytic cells) isolated from the subject at a second time point, a population of circulating vesicles isolated from the subject at a second time point, and a population of circulating diseased cells isolated from the subject at a second time point;

determining a fourth profile of at least one of the one or more markers from a second control comprising a component selected from the group consisting of: a population of phagocytic cells having a DNA content of 2n (=2n phagocytic cells) isolated from the subject at a second time point, a population of non-phagocytic cells isolated from the subject at a second time point, and a population of control cells isolated from the subject, wherein the control cells are substantially free of cells affected by the disease or condition, at a second time point;

identifying a difference between the third and fourth profiles; and

c) identifying a difference between the difference identified in a) and the difference identified in b) wherein the identified difference in c) is indicative of the progression or regression of said disease or condition in the subject.

6. A method for identifying a compound capable of ameliorating or treating a disease or condition in a subject comprising:

a) determining a first profile of one or more markers of the disease or condition from a first sample comprising two or more different components selected from the group consisting of: a cell-free bodily fluid isolated from the subject before administering the compound to the subject, a population of phagocytic cells isolated from the subject before administering the compound to the subject, a population of phagocytic cells having a DNA content more than 2n (>2n phagocytic cells) isolated from the subject before administering the compound to the subject, a population of circulating vesicles isolated from the subject before administering the compound to the subject, and a population of circulating diseased cells isolated from the subject before administering the compound to the subject;

determining a second profile of at least one of the one or more markers from a first control comprising a component selected from the group consisting of: a population of phagocytic cells having a DNA content of 2n (=2n phagocytic cells) isolated from the subject before administering the compound to the subject, a population of non-phagocytic cells isolated from the subject before administering the compound to the subject, and a population of control cells isolated from the subject, wherein the control cells are substantially free of cells affected by the disease or condition, before administering the compound to the subject;

identifying a difference between the first and second profiles;

b) determining a third profile of one or more markers of the disease or condition from a second sample comprising two or more different components selected from the group consisting of: a cell-free bodily fluid isolated from the subject after administering the compound to the subject, a population of phagocytic cells isolated from the subject after administering the compound to the subject, a population of phagocytic cells having a DNA content more than 2n (>2n phagocytic cells) isolated from the subject after administering the compound to the subject, a population of circulating vesicles isolated from the subject after administering the compound to the subject, and a population of circulating diseased cells isolated from the subject after administering the compound to the subject;

determining a fourth profile of at least one of the one or more markers from a second control comprising a component selected from the group consisting of: a population of phagocytic cells having a DNA content of 2n (=2n phagocytic cells) isolated from the subject after administering the compound to the subject, a population of non-phagocytic cells isolated from the subject after administering the compound to the subject, and a population of control cells isolated from the subject, wherein the control cells are substantially free of cells affected by the disease or condition, after administering the compound to the subject;

identifying a difference between the third and fourth profiles; and

c) identifying a difference between the difference identified in a) and the difference identified in b) wherein the identified difference in c) indicates that the compound is capable of ameliorating or treating said disease or condition in the subject.

7. A method for assessing the efficacy of a treatment for a disease or condition in a subject comprising:

a) determining a first profile of one or more markers of the disease or condition from a first sample comprising two or more different components selected from the group consisting of: a cell-free bodily fluid isolated from the subject before the treatment, a population of phagocytic cells isolated from the subject before the treatment, a population of phagocytic cells having a DNA content more than 2n (>2n phagocytic cells) isolated from the subject before the treatment, a population of circulating vesicles isolated from the subject before the treatment, and a population of circulating diseased cells isolated from the subject before the treatment;

b) determining a second profile of one or more markers of the disease or condition from a second sample comprising two or more different components selected from the group consisting of: a cell-free bodily fluid isolated from the subject after the treatment, a population of phagocytic cells isolated from the subject after the treatment, a population of phagocytic cells having a DNA content more than 2n (>2n phagocytic cells) isolated from the subject after the treatment, a population of circulating vesicles isolated from the subject after the treatment, and a population of circulating diseased cells isolated from the subject; and

c) identifying a difference between the first profile and the second profile, wherein the identified difference is indicative of the efficacy of the treatment for said disease or condition in the subject.

8. A method for monitoring the progression or regression of a disease or condition in a subject comprising:

a) determining a first profile of one or more markers of the disease or condition from a first sample comprising two or more different components selected

from the group consisting of: a cell-free bodily fluid isolated from the subject at a first time point, a population of phagocytic cells isolated from the subject at a first time point, a population of phagocytic cells having a DNA content more than 2n (>2n phagocytic cells) isolated from the subject at a first time point, a population of circulating vesicles isolated from the subject at a first time point, and a population of circulating diseased cells isolated from the subject at a first time point;

b) determining a second profile of one or more markers of the disease or condition from a second sample comprising two or more different components selected from the group consisting of: a cell-free bodily fluid isolated from the subject at a second time point, a population of phagocytic cells isolated from the subject at a second time point, a population of phagocytic cells having a DNA content more than 2n (>2n phagocytic cells) isolated from the subject at a second time point, a population of circulating vesicles isolated from the subject at a second time point, and a population of circulating diseased cells isolated from the subject at a second time point; and

c) identifying a difference between the first profile and the second profile, wherein the identified difference is indicative of the progression or regression of said disease or condition in the subject.

9. A method for identifying a compound capable of ameliorating or treating a disease or condition in a subject comprising:

a) determining a first profile of one or more markers of the disease or condition from a first sample comprising two or more different components selected from the group consisting of: a cell-free bodily fluid isolated from the subject before administering the compound to the subject, a population of phagocytic cells isolated from the subject before administering the compound to the subject, a population of phagocytic cells having a DNA content more than 2n (>2n phagocytic cells) isolated from the subject before administering the compound to the subject, a population of circulating vesicles isolated from the subject before administering the compound to the subject, and a population of circulating diseased cells isolated from the subject before administering the compound to the subject;

b) determining a second profile of one or more markers of the disease or condition from a second sample comprising two or more different components selected from the group consisting of: a cell-free bodily fluid isolated from the subject after administering the compound to the subject, a population of phagocytic cells isolated from the subject after administering the compound to the subject, a population of phagocytic cells having a DNA content more than 2n (>2n phagocytic cells) isolated from the subject after administering the compound to the subject, a population of circulating vesicles isolated from the subject after administering the compound to the subject, and a population of circulating diseased cells isolated from the subject after administering the compound to the subject; and

c) identifying a difference between the first profile and the second profile, wherein the identified difference indicates that the compound is capable of ameliorating or treating said disease or condition in the subject.

10. A method for diagnosing or aiding in the diagnosis of a disease or condition in a subject comprising:

a) determining a first profile of one or more markers of the disease or condition from a sample comprising components selected from the group consisting of: an analyte isolated from a cell-free bodily fluid isolated from the subject, an analyte isolated from a population of phagocytic cells isolated from the subject, an analyte isolated from a population of phagocytic cells having a DNA content more than 2n (>2n phagocytic cells) isolated from the subject, an analyte isolated from a population of circulating vesicles isolated from the subject, and an analyte isolated from a population of circulating diseased cells isolated from the subject;

b) determining a second profile of at least one of the one or more markers from a control comprising a component selected from the group consisting of: an analyte isolated from a population of phagocytic cells having a DNA content of 2n (=2n phagocytic cells) isolated from the subject, an analyte isolated from a population of non-phagocytic cells isolated from the subject, and an analyte isolated from a population of control cells isolated from the subject, wherein the control cells are substantially free of cells affected by the disease or condition; and

c) identifying a difference between the first and second profiles, wherein the difference is indicative of the presence of said disease or condition in the subject.

11. A method for assessing the risk of developing a disease or condition in a subject comprising:

a) determining a first profile of one or more markers of the disease or condition from a sample comprising two or more different components selected from the group consisting of: an analyte isolated from a cell-free bodily fluid isolated from the subject, an analyte isolated from a population of phagocytic cells isolated from the subject, an analyte isolated from a population of phagocytic cells having a DNA content more than 2n (>2n phagocytic cells) isolated from the subject, an analyte isolated from a population of circulating vesicles isolated from the subject, and an analyte isolated from a population of circulating diseased cells isolated from the subject;

b) determining a second profile of at least one of the one or more markers from a control comprising a component selected from the group consisting of: an analyte isolated from a population of phagocytic cells having a DNA content of 2n (=2n phagocytic cells) isolated from the subject, an analyte isolated from a population of non-phagocytic cells isolated from the subject, and an analyte isolated from a population of control cells isolated from the subject, wherein the control cells are substantially free of cells affected by the disease or condition; and

c) identifying a difference between the first and second profiles, wherein the difference is indicative of the risk of developing said disease or condition in the subject.

12. A method for prognosing or aiding in the prognosis of a disease or condition in a subject comprising:

a) determining a first profile of one or more markers of the disease or condition from a sample comprising two or more different components selected from the group consisting of: an analyte isolated from a cell-free bodily fluid isolated from the subject, an analyte isolated from a population of phagocytic cells isolated from the

subject, an analyte isolated from a population of phagocytic cells having a DNA content more than 2n (>2n phagocytic cells) isolated from the subject, an analyte isolated from a population of circulating vesicles isolated from the subject, and an analyte isolated from a population of circulating diseased cells isolated from the subject;

b) determining a second profile of at least one of the one or more markers from a control comprising a component selected from the group consisting of: an analyte isolated from a population of phagocytic cells having a DNA content of 2n (=2n phagocytic cells) isolated from the subject, an analyte isolated from a population of non-phagocytic cells isolated from the subject, and an analyte isolated from a population of control cells isolated from the subject, wherein the control cells are substantially free of cells affected by the disease or condition; and

c) identifying a difference between the first and second profiles, wherein the difference is indicative of the prognosis of said disease or condition in the subject.

13. A method for assessing the efficacy of a treatment for a disease or condition in a subject comprising:

a) determining a first profile of one or more markers of the disease or condition from a first sample comprising two or more different components selected from the group consisting of: an analyte isolated from a cell-free bodily fluid isolated from the subject before the treatment, an analyte isolated from a population of phagocytic cells isolated from the subject before the treatment, an analyte isolated from a population of phagocytic cells having a DNA content more than 2n (>2n phagocytic cells) isolated from the subject before the treatment, an analyte isolated from a population of circulating vesicles isolated from the subject before the treatment, and an analyte isolated from a population of circulating diseased cells isolated from the subject before the treatment;

determining a second profile of at least one of the one or more markers from a first control comprising a component selected from the group consisting of: an analyte isolated from a population of phagocytic cells having a DNA content of 2n

(=2n phagocytic cells) isolated from the subject before the treatment, an analyte isolated from a population of non-phagocytic cells isolated from the subject before the treatment, and an analyte isolated from a population of control cells isolated from the subject, wherein the control cells are substantially free of cells affected by the disease or condition, before the treatment;

identifying a difference between the first and second profiles;

b) determining a third profile of one or more markers of the disease or condition from a second sample comprising two or more different components selected from the group consisting of: an analyte isolated from a cell-free bodily fluid isolated from the subject after the treatment, an analyte isolated from a population of phagocytic cells isolated from the subject after the treatment, an analyte isolated from a population of phagocytic cells having a DNA content more than 2n (>2n phagocytic cells) isolated from the subject after the treatment, an analyte isolated from a population of circulating vesicles isolated from the subject after the treatment, and an analyte isolated from a population of circulating diseased cells isolated from the subject after the treatment;

determining a fourth profile of at least one of the one or more markers from a second control comprising a component selected from the group consisting of: an analyte isolated from a population of phagocytic cells having a DNA content of 2n (=2n phagocytic cells) isolated from the subject after the treatment, an analyte isolated from a population of non-phagocytic cells isolated from the subject after the treatment, and an analyte isolated from a population of control cells isolated from the subject, wherein the control cells are substantially free of cells affected by the disease or condition, after the treatment;

identifying a difference between the third and fourth profiles; and

c) identifying a difference between the difference identified in a) and the difference identified in b) wherein the identified difference in c) is indicative of the efficacy of the treatment for said disease or condition in the subject.

14. A method for monitoring the progression or regression of a disease or condition in a subject comprising:

a) determining a first profile of one or more markers of the disease or condition from a first sample comprising two or more different components selected from the group consisting of: an analyte isolated from a cell-free bodily fluid isolated from the subject at a first time point, an analyte isolated from a population of phagocytic cells isolated from the subject at a first time point, an analyte isolated from a population of phagocytic cells having a DNA content more than 2n (>2n phagocytic cells) isolated from the subject at a first time point, an analyte isolated from a population of circulating vesicles isolated from the subject at a first time point, and an analyte isolated from a population of circulating diseased cells isolated from the subject at a first time point;

determining a second profile of at least one of the one or more markers from a first control comprising a component selected from the group consisting of: an analyte isolated from a population of phagocytic cells having a DNA content of 2n (=2n phagocytic cells) isolated from the subject at a first time point, an analyte isolated from a population of non-phagocytic cells isolated from the subject at a first time point, and an analyte isolated from a population of control cells isolated from the subject, wherein the control cells are substantially free of cells affected by the disease or condition, at a first time point;

identifying a difference between the first and second profiles;

b) determining a third profile of one or more markers of the disease or condition from a second sample comprising two or more different components selected from the group consisting of: an analyte isolated from a cell-free bodily fluid isolated from the subject at a second time point, an analyte isolated from a population of phagocytic cells isolated from the subject at a second time point, an analyte isolated from a population of phagocytic cells having a DNA content more than 2n (>2n phagocytic cells) isolated from the subject at a second time point, an analyte isolated from a population of circulating vesicles isolated from the subject at a second

time point, and an analyte isolated from a population of circulating diseased cells isolated from the subject at a second time point;

determining a fourth profile of at least one of the one or more markers from a second control comprising a component selected from the group consisting of: an analyte isolated from a population of phagocytic cells having a DNA content of 2n (=2n phagocytic cells) isolated from the subject at a second time point, an analyte isolated from a population of non-phagocytic cells isolated from the subject at a second time point, and an analyte isolated from a population of control cells isolated from the subject, wherein the control cells are substantially free of cells affected by the disease or condition, at a second time point;

identifying a difference between the third and fourth profiles; and

c) identifying a difference between the difference identified in a) and the difference identified in b) wherein the identified difference in c) is indicative of the progression or regression of said disease or condition in the subject.

15. A method for identifying a compound capable of ameliorating or treating a disease or condition in a subject comprising:

a) determining a first profile of one or more markers of the disease or condition from a first sample comprising two or more different components selected from the group consisting of: an analyte isolated from a cell-free bodily fluid isolated from the subject before administering the compound to the subject, an analyte isolated from a population of phagocytic cells isolated from the subject before administering the compound to the subject, an analyte isolated from a population of phagocytic cells having a DNA content more than 2n (>2n phagocytic cells) isolated from the subject before administering the compound to the subject, an analyte isolated from a population of circulating vesicles isolated from the subject before administering the compound to the subject, and an analyte isolated from a population of circulating diseased cells isolated from the subject before administering the compound to the subject;

determining a second profile of at least one of the one or more markers from a first control comprising a component selected from the group consisting of: an analyte isolated from a population of phagocytic cells having a DNA content of 2n (=2n phagocytic cells) isolated from the subject before administering the compound to the subject, an analyte isolated from a population of non-phagocytic cells isolated from the subject before administering the compound to the subject, and an analyte isolated from a population of control cells isolated from the subject, wherein the control cells are substantially free of cells affected by the disease or condition, before administering the compound to the subject;

identifying a difference between the first and second profiles;

b) determining a third profile of one or more markers of the disease or condition from a second sample comprising two or more different components selected from the group consisting of: an analyte isolated from a cell-free bodily fluid isolated from the subject after administering the compound to the subject, an analyte isolated from a population of phagocytic cells isolated from the subject after administering the compound to the subject, an analyte isolated from a population of phagocytic cells having a DNA content more than 2n (>2n phagocytic cells) isolated from the subject after administering the compound to the subject, an analyte isolated from a population of circulating vesicles isolated from the subject after administering the compound to the subject, and an analyte isolated from a population of circulating diseased cells isolated from the subject after administering the compound to the subject;

determining a fourth profile of at least one of the one or more markers from a second control comprising a component selected from the group consisting of: an analyte isolated from a population of phagocytic cells having a DNA content of 2n (=2n phagocytic cells) isolated from the subject after administering the compound to the subject, an analyte isolated from a population of non-phagocytic cells isolated from the subject after administering the compound to the subject, and an analyte isolated from a population of control cells isolated from the subject, wherein the

control cells are substantially free of cells affected by the disease or condition, after administering the compound to the subject;

identifying a difference between the third and fourth profiles; and

c) identifying a difference between the difference identified in a) and the difference identified in b) wherein the identified difference in c) indicates that the compound is capable of ameliorating or treating said disease or condition in the subject.

16. A method for assessing the efficacy of a treatment for a disease or condition in a subject comprising:

a) determining a first profile of one or more markers of the disease or condition from a first sample comprising two or more different components selected from the group consisting of: an analyte isolated from a cell-free bodily fluid isolated from the subject before the treatment, a population of phagocytic cells isolated from the subject before the treatment, an analyte isolated from a population of phagocytic cells having a DNA content more than 2n (>2n phagocytic cells) isolated from the subject before the treatment, an analyte isolated from a population of circulating vesicles isolated from the subject before the treatment, and an analyte isolated from a population of circulating diseased cells isolated from the subject before the treatment;

b) determining a second profile of one or more markers of the disease or condition from a second sample comprising two or more different components selected from the group consisting of: an analyte isolated from a cell-free bodily fluid isolated from the subject after the treatment, an analyte isolated from a population of phagocytic cells isolated from the subject after the treatment, an analyte isolated from a population of phagocytic cells having a DNA content more than 2n (>2n phagocytic cells) isolated from the subject after the treatment, an analyte isolated from a population of circulating vesicles isolated from the subject after the treatment, and an analyte isolated from a population of control cells isolated from the subject, wherein the control cells are substantially free of cells affected by the disease or condition; and

c) identifying a difference between the first profile and the second profile, wherein the identified difference is indicative of the efficacy of the treatment for said disease or condition in the subject.

17. A method for monitoring the progression or regression of a disease or condition in a subject comprising:

a) determining a first profile of one or more markers of the disease or condition from a first sample comprising two or more different components selected from the group consisting of: an analyte isolated from a cell-free bodily fluid isolated from the subject at a first time point, an analyte isolated from a population of phagocytic cells isolated from the subject at a first time point, an analyte isolated from a population of phagocytic cells having a DNA content more than 2n (>2n phagocytic cells) isolated from the subject at a first time point, an analyte isolated from a population of circulating vesicles isolated from the subject at a first time point, and an analyte isolated from a population of circulating diseased cells isolated from the subject at a first time point;

b) determining a second profile of one or more markers of the disease or condition from a second sample comprising two or more different components selected from the group consisting of: an analyte isolated from a cell-free bodily fluid isolated from the subject at a second time point, an analyte isolated from a population of phagocytic cells isolated from the subject at a second time point, an analyte isolated from a population of phagocytic cells having a DNA content more than 2n (>2n phagocytic cells) isolated from the subject at a second time point, an analyte isolated from a population of circulating vesicles isolated from the subject at a second time point, and an analyte isolated from a population of circulating diseased cells isolated from the subject at a second time point; and

c) identifying a difference between the first profile and the second profile, wherein the identified difference is indicative of the progression or regression of said disease or condition in the subject.

18. A method for identifying a compound capable of ameliorating or treating a disease or condition in a subject comprising:

a) determining a first profile of one or more markers of the disease or condition from a first sample comprising two or more different components selected from the group consisting of: an analyte isolated from a cell-free bodily fluid isolated from the subject before administering the compound to the subject, an analyte isolated from a population of phagocytic cells isolated from the subject before administering the compound to the subject, an analyte isolated from a population of phagocytic cells having a DNA content more than 2n (>2n phagocytic cells) isolated from the subject before administering the compound to the subject, an analyte isolated from a population of circulating vesicles isolated from the subject before administering the compound to the subject, and an analyte isolated from a population of circulating diseased cells isolated from the subject before administering the compound to the subject;

b) determining a second profile of one or more markers of the disease or condition from a second sample comprising two or more different components selected from the group consisting of: an analyte isolated from a cell-free bodily fluid isolated from the subject after administering the compound to the subject, an analyte isolated from a population of phagocytic cells isolated from the subject after administering the compound to the subject, an analyte isolated from a population of phagocytic cells having a DNA content more than 2n (>2n phagocytic cells) isolated from the subject after administering the compound to the subject, an analyte isolated from a population of circulating vesicles isolated from the subject after administering the compound to the subject, and an analyte isolated from a population of circulating diseased cells isolated from the subject after administering the compound to the subject; and

c) identifying a difference between the first profile and the second profile, wherein the identified difference indicates that the compound is capable of ameliorating or treating said disease or condition in the subject.

19. A method for diagnosing or aiding in the diagnosis of a disease or condition in a subject comprising:

a) determining a first profile of one or more markers of the disease or condition from a sample comprising two or more different components selected from the group consisting of: a cell-free bodily fluid isolated from the subject, a population of phagocytic cells isolated from the subject, a population of phagocytic cells having a DNA content more than 2n (>2n phagocytic cells) isolated from the subject, a population of circulating vesicles isolated from the subject, and a population of circulating diseased cells isolated from the subject; and

b) identifying a difference between the first profile and a second profile of at least one of the one or more markers from a repository of said markers of said disease or condition, wherein the difference is indicative of the presence of said disease or condition in the subject.

20. A method for assessing the risk of developing a disease or condition in a subject comprising:

a) determining a first profile of one or more markers of the disease or condition from a sample comprising two or more different components selected from the group consisting of: a cell-free bodily fluid isolated from the subject, a population of phagocytic cells isolated from the subject, a population of phagocytic cells having a DNA content more than 2n (>2n phagocytic cells) isolated from the subject, a population of circulating vesicles isolated from the subject, and a population of circulating diseased cells isolated from the subject; and

b) identifying a difference between the first profile and a second profile of at least one of the one or more markers from a repository of said markers of said disease or condition , wherein the difference is indicative of the risk of developing said disease or condition in the subject.

21. A method for prognosing or aiding in the prognosis of a disease or condition in a subject comprising:

a) determining a first profile of one or more markers of the disease or condition from a sample comprising two or more different components selected from the group consisting of: a cell-free bodily fluid isolated from the subject, a population of phagocytic cells isolated from the subject, a population of phagocytic cells having a DNA content more than 2n (>2n phagocytic cells) isolated from the subject, a population of circulating vesicles isolated from the subject, and a population of circulating diseased cells isolated from the subject; and

b) identifying a difference between the first profile and a second profile of at least one of the one or more markers from a repository of said markers of said disease or condition, wherein the difference is indicative of the prognosis of said disease or condition in the subject.

22. A method for diagnosing or aiding in the diagnosis of a disease or condition in a subject comprising:

a) determining a first profile of one or more markers of the disease or condition from a sample comprising components selected from the group consisting of: an analyte isolated from a cell-free bodily fluid isolated from the subject, an analyte isolated from a population of phagocytic cells isolated from the subject, an analyte isolated from a population of phagocytic cells having a DNA content more than 2n (>2n phagocytic cells) isolated from the subject, an analyte isolated from a population of circulating vesicles isolated from the subject, and an analyte isolated from a population of circulating diseased cells isolated from the subject; and

b) identifying a difference between the first profile and a second profile of at least one of the one or more markers from a repository of said markers of said disease or condition , wherein the difference is indicative of the presence of said disease or condition in the subject.

23. A method for assessing the risk of developing a disease or condition in a subject comprising:

a) determining a first profile of one or more markers of the disease or condition from a sample comprising two or more different components selected from

the group consisting of: an analyte isolated from a cell-free bodily fluid isolated from the subject, an analyte isolated from a population of phagocytic cells isolated from the subject, an analyte isolated from a population of phagocytic cells having a DNA content more than 2n (>2n phagocytic cells) isolated from the subject, an analyte isolated from a population of circulating vesicles isolated from the subject, and an analyte isolated from a population of circulating diseased cells isolated from the subject; and

b) identifying a difference between the first profile and a second profile of at least one of the one or more markers from a repository of said markers of said disease or condition , wherein the difference is indicative of the risk of developing said disease or condition in the subject.

24. A method for prognosing or aiding in the prognosis of a disease or condition in a subject comprising:

a) determining a first profile of one or more markers of the disease or condition from a sample comprising two or more different components selected from the group consisting of: an analyte isolated from a cell-free bodily fluid isolated from the subject, an analyte isolated from a population of phagocytic cells isolated from the subject, an analyte isolated from a population of phagocytic cells having a DNA content more than 2n (>2n phagocytic cells) isolated from the subject, an analyte isolated from a population of circulating vesicles isolated from the subject, and a population of circulating diseased cells isolated from the subject; and

b) identifying a difference between the first profile and a second profile of at least one of the one or more markers from a repository of said markers of said disease or condition , wherein the difference is indicative of the prognosis of said disease or condition in the subject.

25. The method of any one of embodiments 1-3 and 10-12, wherein at least one of the one or more markers is up-regulated or activated in the sample compared to the control.

26. The method of any one of embodiments 1-3 and 10-12, wherein at least one of the one or more markers is down-regulated or inhibited in the sample compared to the control.

27. The method of any one of embodiments 4-6 and 13-15, wherein at least one of the one or more markers is up-regulated or activated in the first sample compared to the first control or in the second sample compared to the second control.

28. The method of any one of embodiments 4-6 and 13-15, wherein at least one of the one or more markers is down-regulated or inhibited in the first sample compared to the first control or in the second sample compared to the second control.

29. The method of any one of embodiments 7-9 and 16-18, wherein at least one of the one or more markers is up-regulated or activated in the first sample compared to the second sample.

30. The method of any one of embodiments 7-9 and 16-18, wherein at least one of the one or more markers is down-regulated or inhibited in the first sample compared to the second sample.

31. The method of any one of embodiments 19-24, wherein at least one of the one or more markers is up-regulated or activated in the sample compared to the repository.

32. The method of any one of embodiments 19-24, wherein at least one of the one or more markers is down-regulated or inhibited in the sample compared to the repository.

33. The method of any one of embodiments 1-32, wherein the first profile or the second profile comprises the absence of at least one of the one or more markers of said disease or condition.

34. The method of any one of embodiments 4-6 and 13-15, wherein the third profile or the fourth profile comprises the absence of at least one of the one or more markers of said disease or condition.

35. The method of any one of embodiments 1-9 and 19-21, wherein, when said method comprises circulating diseased cells, control cells not affected by the disease or condition, phagocytic cells, >2n phagocytic cells, =2n phagocytic cells, or non-phagocytic cells, the method further comprises lysing the circulating diseased cells, control cells not affected by the disease or condition, phagocytic cells, >2n phagocytic cells, =2n phagocytic cells, or non-phagocytic cells.

36. The method of any one of embodiments 1-9, 19-21, and 35 wherein, when said method comprises circulating diseased cells, control cells not affected by the disease or condition, phagocytic cells, >2n phagocytic cells, =2n phagocytic cells, or non-phagocytic cells, the method further comprises extracting at least some of the cellular contents from the circulating diseased cells, control cells not affected by the disease or condition, phagocytic cells, >2n phagocytic cells, =2n phagocytic cells, or non-phagocytic cells.

37. The method of any one of embodiments 1-9 and 19-21, 35, and 36, wherein, when said method comprises the cell-free bodily fluid, the method further comprises extracting the one or more markers from the cell-free bodily fluid.

38. The method of any one of embodiments 1-9 and 19-21, wherein, when said method comprises the cell-free bodily fluid, the cell-free bodily fluid comprises a transrcnal nuclcic acid.

39. The method of any one of embodiments 1-38, wherein at least one of the one or more markers of said disease or condition is present in the circulating diseased cells, cell-free bodily fluid sample, phagocytic cells, or >2n phagocytic cells.

40. The method of any one of embodiments 1-39, wherein at least one of the one or more markers of said disease or condition is not present in the circulating diseased cells, cell-free bodily fluid sample, phagocytic cells, or >2n phagocytic cells.

41. The method of any one of embodiments 1-40, wherein one or more of the circulating diseased cells, control cells not affected by the disease or condition, phagocytic cells, >2n phagocytic cells, =2n phagocytic cells, or non-phagocytic cells are enucleated.

42. The method of embodiment 41, wherein the cells are enucleated using physical removal, chemical treatments, photoablation, or ultraviolet irradiation.

43. The method of embodiment 42, wherein the physical removal uses a micronccdlc, optical twcczcr or aspiration.

44. The method of any one of embodiments 19-24, wherein the repository is obtained by data mining.

45. The method of any one of embodiments 1-44, wherein the phagocytic cells, >2n phagocytic cells, or =2n phagocytic cells are neutrophils, macrophages, monocytes, dendritic cells, foam cells, mast cells, eosinophils, keratinocytes, or mixtures thereof.

46. The method of any one of embodiments 1-45, wherein the non-phagocytic cells are T cells, B cells, null cells, basophils, or mixtures thereof.

47. The method of any one of embodiments 1-46, wherein the circulating diseased cells are blood cells, tumor cells, lymphoma cells, fetal cells, apoptotic cells, epithelia cells, endothelial cells, stem cells, progenitor cells, mesenchymal cells, osteoblast cells, osteocytes, hematopoietic stem cells, foam cells, adipose cells, transcervical cells, circulating cardiocytes, circulating fibrocytes, circulating myocytes, circulating cells from kidney, circulating cells from gastrointestinal tract, circulating cells from lung, circulating cells from reproductive organs, circulating cells from central nervous system, circulating hepatic cells, circulating cells from spleen, circulating cells from thymus, circulating cells from thyroid , circulating cells from an endocrine gland, circulating cells from parathyroid, circulating cells from pituitary, circulating cells from adrenal gland, circulating cells from islets of Langerhans, circulating cells from pancreas, circulating cells from hypothalamus, circulating cells from prostate tissues, circulating cells from breast tissues, circulating cells from circulating retinal cells, circulating ophthalmic cells, circulating auditory cells, circulating epidermal cells, circulating cells from the urinary tract, or mixtures thereof.

48. The method of any one of embodiments 1-47, wherein the control cells are normal cells.

49. The method of any one of embodiments 1-48, wherein the control cells are circulating cells.

50. The method of any one of embodiments 1-49, wherein the circulating vesicles are selected from the group consisting of circulating microvesicles, apoptotic bodies, micro-particles, membrane-bound vesicles, multivesicular bodies, nanovesicles, microparticles, and ARRDC-1 mediated microvesicles (ARMM).

51. The method of embodiment 50, wherein the circulating microvesicles are exosomes or urinary exosomes.

52. The method of any one of embodiments 1-51, wherein the cell-free bodily fluid is separated from a bodily fluid.

53. The method of embodiment 52, wherein the bodily fluid is blood, urine, stool, saliva, lymph fluid, cerebrospinal fluid, synovial fluid, cystic fluid, ascites, pleural effusion, fluid obtained from a pregnant woman in the first trimester, fluid obtained from a pregnant woman in the second trimester, fluid obtained from a pregnant woman in the third trimester, maternal blood, amniotic fluid, chorionic villus sample, fluid from a preimplantation embryo, maternal urine, maternal saliva, placental sample, fetal blood, lavage and cervical vaginal fluid, interstitial fluid, buccal swab sample, sputum, bronchial lavage, Pap smear sample, or ocular fluid.

54. The method of any one of embodiments 52 and 53, wherein the cell-free bodily fluid is separated by filtration, centrifugation, flow cytometry, fluorescence activated cell sorting, gradient-based centrifugation, elution, microfluidics, magnetic separation technique, fluorescent-magnetic separation technique, nanostructure, quantum dots, high throughput microscope-based platforms, or a combination thereof.

55. The method of any one of embodiments 52-54, wherein the cell-free bodily fluid is separated by using a substance present in the sample.

56. The method of embodiment 55, wherein the substance is a product of a marker of said disease or condition.

57. The method of any one of embodiments 1-56, wherein the cell-free bodily fluid is plasma or serum.

58. The method of any one of embodiments 1-57, wherein the phagocytic cells, >2n phagocytic cells, =2n phagocytic cells, or non-phagocytic cells are isolated from a bodily fluid, tissues, or cells of the subject.

59. The method of embodiment 58, wherein the bodily fluid sample is blood, urine, stool, saliva, lymph fluid, cerebrospinal fluid, synovial fluid, cystic fluid, ascites, pleural effusion, fluid obtained from a pregnant woman in the first trimester, fluid obtained from a pregnant woman in the second trimester, fluid obtained from a pregnant woman in the third trimester, maternal blood, amniotic fluid, chorionic villus sample, fluid from a preimplantation embryo, maternal urine, maternal saliva, placental sample, fetal blood, lavage and cervical vaginal fluid, interstitial fluid, buccal swab sample, sputum, bronchial lavage, Pap smear sample, or ocular fluid.

60. The method of embodiment 58, wherein the cells are white blood cells.

61. The method of any one of embodiments 58-60, wherein the phagocytic cells, >2n phagocytic cells, =2n phagocytic cells, or non-phagocytic cells are isolated using antibodies.

62. The method of any one of embodiments 58-60, wherein the phagocytic cells, >2n phagocytic cells, =2n phagocytic cells, or non-phagocytic cells are isolated by flow cytometry, fluorescence activated cell sorting, filtration, gradient-based centrifugation, elution, microfluidics, magnetic separation technique, fluorescent-magnetic separation technique, nanostructure, quantum dots, high throughput microscope-based platforms, or a combination thereof.

63. The method of any one of embodiments 1-62, wherein the one or more markers are nucleic acids, proteins, lipids, carbohydrates, metabolites, or combinations thereof.

64. The method of embodiment 63, wherein the nucleic acids are nucleotides, oligonucleotides, DNAs, RNAs, or DNA-RNA hybrids.

65. The method of embodiment 64, wherein the DNAs are double-stranded DNAs, single-stranded DNAs, multi-stranded DNAs, complementary DNAs, genomic DNAs, or non-coding DNAs.

66. The method of embodiment 64, wherein the RNAs are messenger RNAs (mRNAs), microRNAs (miRNAs), small nucleolar RNAs (snoRNAs), ribosomal RNAs (rRNAs), transfer RNAs (tRNAs), small interfering RNAs (siRNAs), heterogeneous nuclear RNAs (hnRNAs), or small hairpin RNAs (shRNAs).

67. The method of embodiment 63, wherein the proteins are amino acids, peptides, enzymes, antigens, antibodies, cytokines, lipoproteins, glycoproteins, or hormones.

68. The method of embodiment 63, wherein the lipids are fatty acids, neutral fats, phosphatides, cholesterol, cholesterol esters, triglycerides, glycolipids, glycerolipids, glycerophospholipids, sphingolipids, sterol lipids, prenol lipids, saccharolipids, polyketides, choline glycerophospholipid, ethanolamine glycerophospholipid, phosphatidylinositol, phosphatidylglycerol, phosphatidylserine, lyso-choline glycerophospholipid, lyso-ethanolamine glycerophospholipid, phosphatidic acid, lyso-phosphatidic acid, sphingomyelin, galactosylceramide, glucosylceramide, free fatty acids, prostaglandins, triacylglycerol, diacylglycerol, monoacylglycerol, acyl-CoA, acylcarnitine, oxysterol, ceramide, cardiolipin, sphingoid base-1-phosphate, shingosine, lyso-sphingomyelin,, gangliosides, plasmalogen, sulfatide, low density lipoproteins (LDLs), very low density lipoproteins (VLDLs), high density lipoproteins (HDLs), sphingoid base-1-phosphates, or derivatives thereof.

69. The method of embodiment 63, wherein the carbohydrates are monosaccharides, disaccharides, polysaccharides, oligosaccharides, or derivatives thereof.

70. The method of embodiment 63, wherein the metabolites are primary metabolites, secondary metabolites, organic metabolites, inorganic metabolites, prostaglandins, hydroxyeicosatetraenoic acids, hydroxyoctadecadienoic acids, steroids, bile acids, vitamins, or derivatives thereof.

71. The method of any one of embodiments 1-3, 10-12, and 19-24, wherein the profile is a nucleic acid profile, a protein profile, a lipid profile, a carbohydrate profile, a metabolite profile, or a combination thereof.

72. The method of embodiment 65, wherein the profile is determined by a qualitative assay, a quantitative assay, or a combination thereof.

73. The method of any one of embodiments 4-9 and 13-18, wherein the first profile or the second profile is a nucleic acid profile, a protein profile, a lipid profile, a carbohydrate profile, a metabolite profile, or a combination thereof.

74. The method of embodiment 65, wherein the first profile or the second profile is determined by a qualitative assay, a quantitative assay, or a combination thereof.

75. The method of any one of embodiments 4-6 and 74, wherein the third profile or the fourth profile is a nucleic acid profile, a protein profile, a lipid profile, a carbohydrate profile, a metabolite profile, or a combination thereof.

76. The method of embodiment 75, wherein the third profile or the fourth profile is determined by a qualitative assay, a quantitative assay, or a combination thereof.

77. The method of any one of embodiments 72, 66, and 76, wherein the quantitative assay uses sequencing, targeted sequencing, single molecule real-time sequencing, electron microscopy-based sequencing, transistor-mediated sequencing, direct sequencing, random shotgun sequencing, Sanger dideoxy termination sequencing, exon sequencing, whole-genome sequencing, sequencing by hybridization, pyrosequencing, capillary electrophoresis, gel electrophoresis, duplex sequencing, cycle sequencing, single-base extension sequencing, solid-phase sequencing, high-throughput sequencing, massively parallel signature sequencing, emulsion PCR, multiplex PCR, co-amplification at lower denaturation temperature-PCR (COLD-PCR), sequencing by reversible dye terminator, paired-end sequencing, near-term sequencing, exonuclease sequencing, sequencing by ligation, short-read sequencing, single-molecule sequencing, sequencing-by-synthesis, real-time sequencing, reverse-terminator sequencing, nanopore sequencing, 454 sequencing, Solexa Genome Analyzer sequencing, SOLiD® sequencing, MS-PET sequencing,

mass spectrometry, matrix assisted laser desorption/ionization-time of flight (MALDI-TOF) mass spectrometry, electrospray ionization (ESI) mass spectrometry, surface-enhanced laser deorption/ionization-time of flight (SELDI-TOF) mass spectrometry, quadrupole-time of flight (Q-TOF) mass spectrometry, atmospheric pressure photoionization mass spectrometry (APPI-MS), Fourier transform mass spectrometry (FTMS), matrix-assisted laser desorption/ionization-Fourier transform-ion cyclotron resonance (MALDI-FT-ICR) mass spectrometry, secondary ion mass spectrometry (SIMS), polymerase chain reaction (PCR) analysis, co-amplification at lower denaturation temperature-PCR (COLD-PCR), multiplex PCR, quantitative PCR, real-time PCR, fluorescence assay, colorimetric assay, chemiluminescent assay, or a combination thereof.

78. The method of embodiment 77, wherein the nucleic acid profile is a genotypic profile, a single nucleotide polymorphism profile, a gene mutation profile, a gene copy number profile, a DNA methylation profile, a DNA acetylation profile, a chromosome dosage profile, a gene expression profile, or a combination thereof.

79. The method of embodiment 77, wherein the nucleic acid profile is determined by polymerase chain reaction (PCR) analysis, sequencing analysis, electrophoretic analysis, restriction fragment length polymorphism (RFLP) analysis, Northern blot analysis, quantitative PCR, reverse-transcriptase-PCR analysis (RT-PCR), allele-specific oligonucleotide hybridization analysis, comparative genomic hybridization, heteroduplex mobility assay (HMA), single strand conformational polymorphism (SSCP), denaturing gradient gel electrophisis (DGGE), RNAase mismatch analysis, mass spectrometry, tandem mass spectrometry, matrix assisted laser desorption/ionization-time of flight (MALDI-TOF) mass spectrometry, electrospray ionization (ESI) mass spectrometry, surface-enhanced laser deorption/ionization-time of flight (SELDI-TOF) mass spectrometry, quadrupole-time of flight (Q-TOF) mass spectrometry, atmospheric pressure photoionization mass spectrometry (APPI-MS), Fourier transform mass spectrometry (FTMS), matrix-assisted laser desorption/ionization-Fourier transform-ion cyclotron resonance (MALDI-FT-ICR) mass spectrometry, secondary ion mass spectrometry (SIMS), surface plasmon resonance, Southern blot analysis, in situ hybridization, fluorescence in situ

hybridization (FISH), chromogenic in situ hybridization (CISH), immunohistochemistry (IHC), microarray, comparative genomic hybridization, karyotyping, multiplex ligation-dependent probe amplification (MLPA), Quantitative Multiplex PCR of Short Fluorescent Fragments (QMPSF), microscopy, methylation specific PCR (MSP) assay, HpaII tiny fragment Enrichment by Ligation-mediated PCR (HELP) assay, radioactive acetate labeling assays, colorimetric DNA acetylation assay, chromatin immunoprecipitation combined with microarray (ChIP-on-chip) assay, restriction landmark genomic scanning, Methylated DNA immunoprecipitation (MeDIP), molecular break light assay for DNA adenine methyltransferase activity, chromatographic separation, methylation-sensitive restriction enzyme analysis, bisulfite-driven conversion of non-methylated cytosine to uracil, co-amplification at lower denaturation temperature-PCR (COLD-PCR), multiplex PCR, methyl-binding PCR analysis, or a combination thereof.

80. The method of embodiment 77, wherein the nucleic acid profile is determined by a sequencing technique selected from the group consisting of targeted sequencing, single molecule real-time sequencing, exon sequencing, electron microscopy-based sequencing, transistor-mediated sequencing, direct sequencing, random shotgun sequencing, Sanger dideoxy termination sequencing, whole-genome sequencing, sequencing by hybridization, pyrosequencing, capillary electrophoresis, gel electrophoresis, duplex sequencing, cycle sequencing, single-base extension sequencing, solid-phase sequencing, high-throughput sequencing, massively parallel signature sequencing, emulsion PCR, co-amplification at lower denaturation temperature-PCR (COLD-PCR), multiplex PCR, sequencing by reversible dye terminator, paired-end sequencing, near-term sequencing, exonuclease sequencing, sequencing by ligation, short-read sequencing, single-molecule sequencing, sequencing-by-synthesis, real-time sequencing, reverse-terminator sequencing, nanopore sequencing, 454 sequencing, Solexa Genome Analyzer sequencing, SOLiD® sequencing, MS-PET sequencing, mass spectrometry, and a combination thereof.

81. The method of embodiment 77, wherein the protein profile is a protein expression profile, a protein activation profile, or a combination thereof.

82. The method of embodiment 77, wherein the protein profile is determined by an immunohistochemistry assay, an enzyme-linked immunosorbent assay (ELISA), in situ hybridization, chromatography, liquid chromatography, size exclusion chromatography, high performance liquid chromatography (HPLC), gas chromatography, mass spectrometry, tandem mass spectrometry, matrix assisted laser desorption/ionization-time of flight (MALDI-TOF) mass spectrometry, electrospray ionization (ESI) mass spectrometry, surface-enhanced laser deorption/ionization-time of flight (SELDI-TOF) mass spectrometry, quadrupole-time of flight (Q-TOF) mass spectrometry, atmospheric pressure photoionization mass spectrometry (APPI-MS), Fourier transform mass spectrometry (FTMS), matrix-assisted laser desorption/ionization-Fourier transform-ion cyclotron resonance (MALDI-FT-ICR) mass spectrometry, secondary ion mass spectrometry (SIMS), radioimmunoassays, microscopy, microfluidic chip-based assays, surface plasmon resonance, sequencing, Western blotting assay, or a combination thereof.

83. The method of embodiment 77, wherein the protein activation profile comprises determining a phosphorylation state, an ubiquitination state, a myristoylation state, a conformational state, or a combination thereof of the one or more markers.

84. The method of embodiment 77, wherein the lipid profile is determined by chromatography, liquid chromatography, size exclusion chromatography, high performance liquid chromatography (HPLC), gas chromatography, mass spectrometry, tandem mass spectrometry, matrix assisted laser desorption/ionization-time of flight (MALDI-TOF) mass spectrometry, electrospray ionization (ESI) mass spectrometry, surface-enhanced laser deorption/ionization-time of flight (SELDI-TOF) mass spectrometry, quadrupole-time of flight (Q-TOF) mass spectrometry, atmospheric pressure photoionization mass spectrometry (APPI-MS), Fourier transform mass spectrometry (FTMS), matrix-assisted laser desorption/ionization-Fourier transform-ion cyclotron resonance (MALDI-FT-ICR) mass spectrometry, secondary ion mass spectrometry (SIMS), radioimmunoassays, microfluidic chip-based assay, detection of fluorescence, detection of chemiluminescence, or a combination thereof.

85. The method of embodiment 77, wherein the carbohydrate profile is determined by chromatography, liquid chromatography, size exclusion chromatography, high performance anion exchange chromatography with pulsed amperometric detection (HPAEC-PAD), liquid chromatography, gas chromatography, fluorescent assay, mass spectrometry, tandem mass spectrometry, matrix assisted laser desorption/ionization-time of flight (MALDI-TOF) mass spectrometry, electrospray ionization (ESI) mass spectrometry, surface-enhanced laser deorption/ionization-time of flight (SELDI-TOF) mass spectrometry, quadrupole-time of flight (Q-TOF) mass spectrometry, atmospheric pressure photoionization mass spectrometry (APPI-MS), Fourier transform mass spectrometry (FTMS), matrix-assisted laser desorption/ionization-Fourier transform-ion cyclotron resonance (MALDI-FT-ICR) mass spectrometry, secondary ion mass spectrometry (SIMS), radioimmunoassay, microfluidic chip-based assay, detection of fluorescence, detection of chemiluminescence, or a combination thereof.

86. The method of any one of embodiments 1-85, wherein the subject has at least two diseases or conditions.

87. The method of embodiment 86, wherein the subject has at least one prenatal or pregnancy-related disease or condition

88. The method of any one of embodiments 1-87, wherein the subject is a mammal.

89. The method of embodiment 89, wherein the mammal is a human.

90. The method of any one of embodiments 1-89, wherein the disease or condition is a cardiovascular disease or condition, a kidney-associated disease or condition, a prenatal or pregnancy-related disease or condition, a neurological or neuropsychiatric disease or condition, an autoimmune or immune-related disease or condition, a cancer, an infectious disease or condition, a mitochondrial disorder, a respiratory-gastrointestinal tract disease or condition, a reproductive disease or condition, an ophthalmic disease or condition, a musculo-skeletal disease or condition, or a dermal disease or condition.

91. The method of any one of embodiments 1-90, wherein the difference is greater than a 1-fold difference.

92. The method of embodiment 91, wherein the difference is at least 1.05-fold, 1.1-fold, 1.2-fold, 1.3-fold, 1.4-fold, 1.5-fold, 2-fold, 2.5-fold, 3-fold, 4-fold, 5-fold, 6-fold, 7-fold, 8-fold, 9-fold, or 10-fold difference.

93. The method of any one of embodiments 4-6 and 13-15, wherein the second profile and the fourth profile are the same.

94. A method for identifying one or more markers for a disease or condition comprising:

a) determining a first profile of analytes from a sample comprising a cell-free bodily fluid from a subject having said disease or condition and a population of phagocytic cells, or a population of >2n phagocytic cells, from a subject having said disease or condition;

determining a second profile of analytes from a population of =2n phagocytic cells, or a population of non-phagocytic cells, from the subject having said disease or condition;

identifying a set of differences between the first and second profiles, wherein the first set of differences is specific to the first profile relative to the second profile;

b) determining a third profile of analytes from a sample comprising a cell-free bodily fluid from a subject having said disease or condition and a population of phagocytic cells, or a population of >2n phagocytic cells, from a control subject not having said disease or condition;

determining a fourth profile of analytes from a population of =2n phagocytic cells, or a population of non-phagocytic cells, from the control subject not having said disease or condition;

identifying a set of differences between the third and fourth profiles, wherein the second set of differences is specific to the third profile relative to the fourth profile; and

c) identifying one or more analytes specific to the set of differences identified in a) relative to the set of differences identified in b), the identified analytes in c) being markers of said disease or condition.

95. A method for identifying one or more markers for a disease or condition comprising:

a) determining a first profile of analytes from a sample comprising a cell-free bodily fluid from a subject having said disease or condition and a population of phagocytic cells, or a population of >2n phagocytic cells, from a subject having said disease or condition;

b) comparing the first profile to a second profile derived from a repository of analytes from a control subject not having said disease or condition;

c) identifying a set of differences between the first and second profiles, wherein the set of differences is specific to the first profile relative to the second profile; and

d) identifying one or more analytes specific to the set of differences, the identified analytes being markers of said disease or condition.

96. The method of embodiment 95, further comprising:

a) obtaining a fifth profile of analytes from cells or tissues affected by said disease or condition in the subject having said disease or condition;

obtaining a sixth profile of analytes from cells or tissues not affected by said disease or condition in the subject having said disease or condition;

identifying a set of differences between the fifth and sixth profiles, wherein the set of differences is specific to the fifth profile relative to the sixth profile; and

b) identifying at least one of the one or more markers of c) present in the set of differences identified in d).

97. The method of any one of embodiments 90-96, further comprising extracting the one or more markers from the cell-free bodily fluid.

98. The method of any one of embodiments 90-97, further comprising lysing the phagocytic cells, the >2n phagocytic cells, the =2n phagocytic cells, or the non-phagocytic cells before a).

99. The method of any one of embodiments 90-98, further comprising extracting at least some of the cellular contents from the phagocytic cells, the >2n phagocytic cells, the =2n phagocytic cells, or the non-phagocytic cells before a).

100. The method of any one of embodiments 90-99, wherein the phagocytic cells or the >2n phagocytic cells comprise viable diseased cells, dead diseased cells, apoptotic diseased cells, circulating tumor cells, infectious agents, fetal cells, trophoblasts, or fragments thereof.

101. The method of any one of embodiments 90-100, wherein at least one of the one or more markers of said disease or condition is present in the cell-free bodily fluid sample, the phagocytic cells, or the >2n phagocytic cells.

102. The method of any one of embodiments 90-101, wherein at least one of the one or more markers of said disease or condition is not present in the cell-free bodily fluid sample, the phagocytic cells, or the >2n phagocytic cells.

103. The method of any one of embodiments 90-102, further comprising comparing the identified difference of c) to a repository of one or more known markers of said disease or condition.

104. The method of embodiment 97, wherein the repository is obtained by data mining.

105. The method of any one of embodiments 90-104, wherein the phagocytic cells, >2n phagocytic cells, or =2n phagocytic cells are neutrophils, macrophages, monocytes, dendritic cells, foam cells, mast cells, eosinophils, keratinocytes, or mixtures thereof.

106. The method of any one of the embodiments 90-105, wherein the non-phagocytic cells are T cells, B cells, null cells, basophils, or mixtures thereof

107. The method of any one of embodiments 90-106, wherein the cell-free bodily fluid is separated from a bodily fluid.

108. The method of embodiment 107, wherein the bodily fluid sample is blood, urine, stool, saliva, lymph fluid, cerebrospinal fluid, synovial fluid, cystic fluid, ascites, pleural effusion, fluid obtained from a pregnant woman in the first trimester, fluid obtained from a pregnant woman in the second trimester, fluid obtained from a pregnant woman in the third trimester, maternal blood, amniotic fluid, chorionic villus sample, fluid from a preimplantation embryo, maternal urine, maternal saliva, placental sample, fetal blood, lavage and cervical vaginal fluid, interstitial fluid, or ocular fluid.

109. The method of any one of embodiments 107 and 108, wherein the cell-free bodily fluid sample is separated by filtration, centrifugation, flow cytometry, fluorescence activated cell sorting, gradient-based centrifugation, elution, microfluidics, magnetic separation technique, fluorescent-magnetic separation technique, nanostructure, quantum dots, high throughput microscope-based platforms, or a combination thereof.

110. The method of embodiment 109, wherein the cell-free bodily fluid sample is separated by using a substance present in the sample.

111. The method of any one of embodiments 90-110, wherein the one or more markers are nucleic acids, proteins, lipids, carbohydrates, metabolites, or combinations thereof.

112. The method of embodiment 111, wherein the nucleic acids are nucleotides, oligonucleotides, DNAs, RNAs, or DNA-RNA hybrids.

113. The method of embodiment 112, wherein the DNAs are double-stranded DNAs, single-stranded DNAs, multi-stranded DNAs, complementary DNAs, genomic DNAs, or non-coding DNAs.

114. The method of embodiment 112, wherein the RNAs are messenger RNAs (mRNAs), microRNAs (miRNAs), small nucleolar RNAs (snoRNAs), ribosomal RNAs (rRNAs), transfer RNAs (tRNAs), small interfering RNAs (siRNAs), heterogeneous nuclear RNAs (hnRNAs), or small hairpin RNAs (shRNAs).

115. The method of embodiment 111, wherein the proteins are amino acids, peptides, enzymes, antigens, antibodies, cytokines, lipoproteins, glycoproteins, or hormones.

116. The method of embodiment 111, wherein the lipids are fatty acids, neutral fats, phosphatides, cholesterol, cholesterol esters, triglycerides, glycolipids, glycerolipids, glycerophospholipids, sphingolipids, sterol lipids, prenol lipids, saccharolipids, polyketides, choline glycerophospholipid, ethanolamine glycerophospholipid, phosphatidylinositol, phosphatidylglycerol, phosphatidylserine, lyso-choline glycerophospholipid, lyso-ethanolamine glycerophospholipid, phosphatidic acid, lyso-phosphatidic acid, sphingomyelin, galactosylceramide, glucosylceramide, free fatty acids, prostaglandins, triacylglycerol, diacylglycerol, monoacylglycerol, acyl-CoA, acylcarnitine, oxysterol, ceramide, cardiolipin, sphingoid base-1-phosphate, shingosine, lyso-sphingomyelin,, gangliosides, plasmalogen, sulfatide, low density lipoproteins (LDLs), very low density lipoproteins (VLDLs), high density lipoproteins (HDLs), sphingoid base-1-phosphates, or derivatives thereof.

117. The method of embodiment 111, wherein the carbohydrates are monosaccharides, disaccharides, polysaccharides, oligosaccharides, or derivatives thereof.

118. The method of embodiment 111, wherein the metabolites are primary metabolites, secondary metabolites, organic metabolites, inorganic metabolites, prostaglandins, hydroxyeicosatetraenoic acids, hydroxyoctadecadienoic acids, steroids, bile acids, vitamins, or derivatives thereof.

119. The method of any one of embodiments 90-118, wherein the profile is a nucleic acid profile, a protein profile, a lipid profile, a carbohydrate profile, a metabolite profile, or a combination thereof.

120. The method of embodiment 119, wherein the profile is determined by a qualitative assay, a quantitative assay, or a combination thereof.

121. The method of embodiment 120, wherein the quantitative assay uses sequencing, targeted sequencing, single molecule real-time sequencing, exon sequencing, electron microscopy-based sequencing, transistor-mediated sequencing, direct sequencing, random shotgun sequencing, Sanger dideoxy termination sequencing, whole-genome sequencing, sequencing by hybridization, pyrosequencing, capillary electrophoresis, gel electrophoresis, duplex sequencing, cycle sequencing, single-base extension sequencing, solid-phase sequencing, high-throughput sequencing, massively parallel signature sequencing, emulsion PCR, co-amplification at lower denaturation temperature-PCR (COLD-PCR), multiplex PCR, sequencing by reversible dye terminator, paired-end sequencing, near-term sequencing, exonuclease sequencing, sequencing by ligation, short-read sequencing, single-molecule sequencing, sequencing-by-synthesis, real-time sequencing, reverse-terminator sequencing, nanopore sequencing, 454 sequencing, Solexa Genome Analyzer sequencing, SOLiD® sequencing, MS-PET sequencing, mass spectrometry, matrix assisted laser desorption/ionization-time of flight (MALDI-TOF) mass spectrometry, electrospray ionization (ESI) mass spectrometry, surface-enhanced laser desorption/ionization-time of flight (SELDI-TOF) mass spectrometry, quadrupole-time of flight (Q-TOF) mass spectrometry, atmospheric pressure photoionization mass

spectrometry (APPI-MS), Fourier transform mass spectrometry (FTMS), matrix-assisted laser desorption/ionization-Fourier transform-ion cyclotron resonance (MALDI-FT-ICR) mass spectrometry, secondary ion mass spectrometry (SIMS), polymerase chain reaction (PCR) analysis, quantitative PCR, real-time PCR, fluorescence assay, colorimetric assay, chemiluminescent assay, or a combination thereof.

122. The method of embodiment 119, wherein the nucleic acid profile is a genotypic profile, a single nucleotide polymorphism profile, a gene mutation profile, a gene copy number profile, a DNA methylation profile, a DNA acetylation profile, a chromosome dosage profile, a gene expression profile, or a combination thereof.

123. The method of embodiment 119, wherein the nucleic acid profile is determined by polymerase chain reaction (PCR) analysis, sequencing analysis, electrophoretic analysis, restriction fragment length polymorphism (RFLP) analysis, Northern blot analysis, quantitative PCR, reverse-transcriptase-PCR analysis (RT-PCR), allele-specific oligonucleotide hybridization analysis, comparative genomic hybridization, heteroduplex mobility assay (HMA), single strand conformational polymorphism (SSCP), denaturing gradient gel electrophisis (DGGE), RNAase mismatch analysis, mass spectrometry, tandem mass spectrometry, matrix assisted laser desorption/ionization-time of flight (MALDI-TOF) mass spectrometry, electrospray ionization (ESI) mass spectrometry, surface-enhanced laser deorption/ionization-time of flight (SELDI-TOF) mass spectrometry, quadrupole-time of flight (Q-TOF) mass spectrometry, atmospheric pressure photoionization mass spectrometry (APPI-MS), Fourier transform mass spectrometry (FTMS), matrix-assisted laser desorption/ionization-Fourier transform-ion cyclotron resonance (MALDI-FT-ICR) mass spectrometry, secondary ion mass spectrometry (SIMS), surface plasmon resonance, Southern blot analysis, in situ hybridization, fluorescence in situ hybridization (FISH), chromogenic in situ hybridization (CISH), immunohistochemistry (IHC), microarray, comparative genomic hybridization, karyotyping, multiplex ligation-dependent probe amplification (MLPA), Quantitative Multiplex PCR of Short Fluorescent Fragments (QMPSF), microscopy, methylation specific PCR (MSP) assay, HpaII tiny fragment Enrichment by Ligation-mediated

PCR (HELP) assay, radioactive acetate labeling assays, colorimetric DNA acetylation assay, chromatin immunoprecipitation combined with microarray (ChIP-on-chip) assay, restriction landmark genomic scanning, Methylated DNA immunoprecipitation (MeDIP), molecular break light assay for DNA adenine methyltransferase activity, chromatographic separation, methylation-sensitive restriction enzyme analysis, bisulfite-driven conversion of non-methylated cytosine to uracil, co-amplification at lower denaturation temperature-PCR (COLD-PCR), multiplex PCR, methyl-binding PCR analysis, or a combination thereof.

124. The method of embodiment 119, wherein the nucleic acid profile is determined by a sequencing technique selected from the group consisting of targeted sequencing, single molecule real-time sequencing, exon sequencing, electron microscopy-based sequencing, transistor-mediated sequencing, direct sequencing, random shotgun sequencing, Sanger dideoxy termination sequencing, whole-genome sequencing, sequencing by hybridization, pyrosequencing, capillary electrophoresis, gel electrophoresis, duplex sequencing, cycle sequencing, single-base extension sequencing, solid-phase sequencing, high-throughput sequencing, massively parallel signature sequencing, emulsion PCR, co-amplification at lower denaturation temperature-PCR (COLD-PCR), multiplex PCR, sequencing by reversible dye terminator, paired-end sequencing, near-term sequencing, exonuclease sequencing, sequencing by ligation, short-read sequencing, single-molecule sequencing, sequencing-by-synthesis, real-time sequencing, reverse-terminator sequencing, nanopore sequencing, 454 sequencing, Solexa Genome Analyzer sequencing, SOLiD® sequencing, MS-PET sequencing, mass spectrometry, and a combination thereof.

125. The method of embodiment 119, wherein the protein profile is a protein expression profile, a protein activation profile, or a combination thereof.

126. The method of embodiment 119, wherein the protein profile is determined by an immunohistochemistry assay, an enzyme-linked immunosorbent assay (ELISA), in situ hybridization, chromatography, liquid chromatography, size exclusion chromatography, high performance liquid chromatography (HPLC), gas

chromatography, mass spectrometry, tandem mass spectrometry, matrix assisted laser desorption/ionization-time of flight (MALDI-TOF) mass spectrometry, electrospray ionization (ESI) mass spectrometry, surface-enhanced laser deorption/ionization-time of flight (SELDI-TOF) mass spectrometry, quadrupole-time of flight (Q-TOF) mass spectrometry, atmospheric pressure photoionization mass spectrometry (APPI-MS), Fourier transform mass spectrometry (FTMS), matrix-assisted laser desorption/ionization-Fourier transform-ion cyclotron resonance (MALDI-FT-ICR) mass spectrometry, secondary ion mass spectrometry (SIMS), radioimmunoassays, microscopy, microfluidic chip-based assays, surface plasmon resonance, sequencing, Western blotting assay, or a combination thereof.

127. The method of embodiment 119, wherein the protein activation profile comprises determining a phosphorylation state, an ubiquitination state, a myristoylation state, a conformational state, or a combination thereof of the one or more markers.

128. The method of embodiment 119, wherein the lipid profile is determined by chromatography, liquid chromatography, size exclusion chromatography, high performance liquid chromatography (HPLC), gas chromatography, mass spectrometry, tandem mass spectrometry, matrix assisted laser desorption/ionization-time of flight (MALDI-TOF) mass spectrometry, electrospray ionization (ESI) mass spectrometry, surface-enhanced laser deorption/ionization-time of flight (SELDI-TOF) mass spectrometry, quadrupole-time of flight (Q-TOF) mass spectrometry, atmospheric pressure photoionization mass spectrometry (APPI-MS), Fourier transform mass spectrometry (FTMS), matrix-assisted laser desorption/ionization-Fourier transform-ion cyclotron resonance (MALDI-FT-ICR) mass spectrometry, secondary ion mass spectrometry (SIMS), radioimmunoassays, microfluidic chip-based assay, detection of fluorescence, detection of chemiluminescence, or a combination thereof.

129. The method of embodiment 119, wherein the carbohydrate profile is determined by chromatography, liquid chromatography, size exclusion chromatography, high performance anion exchange chromatography with pulsed

amperometric detection (HPAEC-PAD), liquid chromatography, gas chromatography, fluorescent assay, mass spectrometry, tandem mass spectrometry, matrix assisted laser desorption/ionization-time of flight (MALDI-TOF) mass spectrometry, electrospray ionization (ESI) mass spectrometry, surface-enhanced laser deorption/ionization-time of flight (SELDI-TOF) mass spectrometry, quadrupole-time of flight (Q-TOF) mass spectrometry, atmospheric pressure photoionization mass spectrometry (APPI-MS), Fourier transform mass spectrometry (FTMS), matrix-assisted laser desorption/ionization-Fourier transform-ion cyclotron resonance (MALDI-FT-ICR) mass spectrometry, secondary ion mass spectrometry (SIMS), radioimmunoassay, microfluidic chip-based assay, detection of fluorescence, detection of chemiluminescence, or a combination thereof.

130. The method of any one of embodiments 90-129, wherein the subject is a mammal

131. The method of embodiment 130, wherein the mammal is a human.

132. The method of any one of embodiments 90-131, wherein the disease or condition is a cardiovascular disease or condition, a kidney-associated disease or condition, a prenatal or pregnancy-related disease or condition, a neurological or neuropsychiatric disease or condition, an autoimmune or immune-related disease or condition, a cancer, an infectious disease or condition, a mitochondrial disorder, a respiratory-gastrointestinal tract disease or condition, a reproductive disease or condition, an ophthalmic disease or condition, a musculo-skeletal disease or condition, or a dermal disease or condition.

133. The method of any one of embodiments 90-132, wherein the difference is greater than a 1-fold difference.

134. The method of embodiment 133, wherein the difference is at least 1.05-fold, 1.1-fold, 1.2-fold, 1.3-fold, 1.4-fold, 1.5-fold, 2-fold, 2.5-fold, 3-fold, 4-fold, 5-fold, 6-fold, 7-fold, 8-fold, 9-fold, or 10-fold difference.

135. The method of any one of embodiments 1-93, further comprising determining at least one diagnostic parameter of said disease or condition.

136. The method of embodiment 135, wherein the diagnostic parameter is determined by physical inspection, visual inspection, biopsy, scanning, histology, radiology, imaging, ultrasound, use of a commercial kit, genetic testing, immunological testing, analysis of bodily fluids, or monitoring neural activity.

137. The method of any one of the embodiments 1-93 and 135-136, wherein the one or more markers comprise at least one gene selected from the group consisting of AKT2, BAK1, EGFR, ERBB2, ETS2, FOS, JUN, MAP2K1, MMP2, PDGFB, RB1, SERPINB2, SNCG, and SPP1.

138. The method of any one of the embodiments 1-93 and 135-137, wherein the one or more markers comprise at least one gene selected from the group consisting of AKT1, AKT2, BAK2, CDC25A, E2F1, EGFR, ERBB2, FOS, JUN, MAP2K1, MMP2, NFKB1, PDGFB, PIK3R1, PNN, RB1, SERPINB2, SERPINB5, SNCG, SPP1, TERT, TIMP3, and TP53.

139. The method of any one of the embodiments 1-93 and 135-138, wherein the one or more markers comprise at least one gene selected from the group consisting of CASP8, CASP9, COL18A1, ETS2, HTATIP2, MMP9, SRC, and TWIST1.

140. The method of any one of the embodiments 1-93 and 135-139, wherein the one or more markers comprise at least one gene selected from the group consisting of AKT1, APAF1, ATM, CDC25A, CDKN1A, ETS2, FOS, IL8, ITGA4, ITGA6, ITGAV, JUN, MAP2K1, NFKBIA, PLAU, PLAUR, RAF1, SERPINB2, SYK, TIMP1, TNF, TNFRSF10B, and TNFRSF1A.

141. The method of any one of the embodiments 1-93 and 135-140, wherein the one or more markers comprise at least one gene selected from the group consisting of ACP2, AK2, AKT3, ARL5B, ATP2B3, BGN, BRAF, BTG2, CAMKK2, CAPG, CAPN12, CPLX2, DENND5A, DNA2, FAM104A, FNIP1, GFRA4, GLUD1, GNAQ, GP1BB, HNRPLL, HOXA2, HPS3, INPP4A, ITGAV, KLHL23, LANCL2, LYPD6, MAPKAPK3, MEF2A (includes, EG:4205), MEF2C, NVL, PCYT1A, PGLYRP4, PLOD1, PPP1CB, PRKAB2, PROS1, PTPRE, RASA4

(includes,EG:10156), RBMS2, RBPJ, STAT5B, THBS1, TRIB1, TRIM2, TSPAN6, and ZDHHC21.

142. The method of any one of the embodiments 1-93 and 135-141, wherein the one or more markers comprise at least one gene selected from the group consisting of B4GALT5, BOP1, CCL2, CCL3, CCL3L1, CCRL2, CD83, CLEC4G, CLIC4, CTSC, CTSO, CXCL10, FCGR3A, FPR3, HBA1, HBB, LRMP, MAP1LC3B2, MS4A4A, MSR1, MYADML, NID1, PF4, PION, RNF217, SAMD9L, SERPING1, and SPARC.

143. The method of any one of the embodiments 1-93 and 135-142, wherein the one or more markers comprise at least one gene selected from the group consisting of ACOT9, AMPD2, ARHGAP15, BATF2, C3AR1, C5orf41, CCL3, CCL3L1, CD63, CHST11, CHSY1, CLEC4G, CTSZ, CXorf21, CYTH4, CYTIP, DLEU2, DNAJA1, DOCK8, DTX3L, DUSP6, EPSTI1, ERF, F2RL1, FYB, GABRB2, GBP5, GLRX, GNB4, ICAM1, IFI35, IFIH1, IFNAR2, IL1R1, IRF1, ITGA5, LAP3, LAPTM5, LCP2, MAP1LC3B, MAP1LC3B2, MICAL2, MT1DP, MT1JP, MT1M, MT2A, MYADML, NEK6, NINJ2, NNMT, NT5C3L, NUB1, PDE4B, PLOD1, PML, PRKCB, PSMB9, RCN3, RGS4, RNASE6, RTP4, SAMD9L, SEL1L, SERPING1, SETX, SIGLEC10, SKIL, SLC7A7, SNORA21, SP100, SP110, SP140, SSFA2, STAT2, STK17B, STK3, TDRD7, TMCC1, TMPRSS11E2, TNFRSF1B, TPM1, TRIM21,TXNDC4, UBE2L6, UBE2W, USP18, VAV1, WARS, WIPF1, and WIPI1.

144. The method of any one of the embodiments 1-93 and 135-143, wherein the one or more markers comprise at least one gene selected from the group consisting of ADAR, ADM, ALAS1, ANKRD22, ARHGAP27, B3GNT5, BCL10, C12orf35, C15orf29, C2orf59, CD177, CEACAM1, CPEB2, DDX58, F2RL1, GDPD3, GNAI3, HIST2H3A, HIST2H3D, HIST2H4A, HMGCR, HSPA6, HSPC159, IL4R, IMPA2, KPNB1, KREMEN1, KRT23, LDLR, LOC100130904, LTB4R, MAEA, MARK2, MBOAT2, MPZL3, N4BP1, NBEAL2, NMI, NPEPPS, PARP14, PGM2, PPIF, PXN, RALBP1, ROD1, RPS6KA1, S100P, SERTAD2, SLC9A1, SLPI, SP110, SPINT1, ST14, TBC1D3, TNFRSF9, TRIM21, UPP1, VPS24, ZBTB34, and ZNF256.

145. The method of any one of the embodiments 1-93 and 135-144, wherein the one or more markers comprise at least one or more of the markers identified by the methods of any one the embodiments 94-134.

146. A kit comprising a plurality of marker detection agents that detect at least one or more of the markers identified by the methods of any one of the embodiments 94-134.

147. A method of treating or preventing a disease or condition in a subject comprising administering to said subject a composition comprising a compound identified by the method of any one of embodiments 6 and 15.

148. The method of any one of the embodiments 1-93 and 135-145, wherein the circulating diseased cells are infected by an infectious agent.

149. The method of embodiment 148, wherein the infectious agent is a virus, bacteria, fungus, parasite, protozoan, infectious protein or microorganism.

150. The method of any one of embodiments 1-93, 135-145, and 148-149, wherein, when said method comprises phagocytic cells or >2n phagocytic cells, said phagocytic cells or >2n phagocytic cells comprise a transfetal nucleic acid.

Brief Description of the Drawings

[0008] Figure 1 depicts a set of potentially informative markers and the relative percentage of reads of associated allelic frequencies in maternal reference samples of CD2-positive peripheral blood cells and combination samples. Along the x-axis are the markers and along the y-axis are the percent reads for the different alleles. Triangle: percent of maternal sample reference allele; circle: percent of maternal sample alternate allele; square: percent of combination sample reference allele; and diamond: percent of combination sample alternate allele.

[0009] Figure 2 depicts a set of potentially informative markers and the relative percentage of reads of associated allelic frequencies in maternal reference samples and combination samples. Along the x-axis are the markers and along the y-axis are the percent reads for the different alleles. Triangle: percent of maternal sample

reference allele; circle: percent of maternal sample alternate allele; square: percent of combination sample reference allele; and diamond: percent of combination sample alternate allele.

[0010] Figure 3 depicts a set of potentially informative markers and the relative percentage of reads of associated allelic frequencies as compared between a circulating cell-free fluid (ccff) sample and a maternal reference sample. Along the x-axis are the markers and along the y-axis are the percent reads for the different alleles. Triangle: reference allele; circle: alternative allele.

[0011] Figure 4 depicts a set of potentially informative markers and the relative percentage of reads of associated allelic frequencies in a T-cell sample. Along the x-axis are the markers and along the y-axis are the percent reads for the different alleles. Triangle: reference allele; circle: alternative allele.

[0012] Figure 5 depicts a set of potentially informative markers and the relative percentage of reads of associated allelic frequencies in a combination sample of a circulating cell-free fluid (ccff) plus 1% monocyte sample. Along the x-axis are the markers and along the y-axis are the percent reads for the different alleles. Triangle: reference allele; circle: alternative allele.

[0013] Figure 6 depicts the emergence of informative markers in a circulating cell-free fluid fraction of the maternal blood compared to the lack of informative markers in the maternal T-cells. Along the x-axis are the markers and along the y-axis are the percent reads for the different alleles.

[0014] Figure 7 depicts the emergence of the informative markers in a combination sample containing cell-free bodily fluid plus monocytes from maternal blood compared to a lack of informative markers in the maternal reference sample containing T-cells. Along the x-axis are the markers and along the y-axis are the percent reads for the different alleles.

[0015] Figure 8 depicts an exemplary informative marker in a maternal only sample and a combination sample. Triangle: reference allele; circle: alternative allele.

[0016] Figure 9 depicts an exemplary informative marker in a maternal only sample and a combination sample. Triangle: reference allele; circle: alternative allele.

[0017] Figure 10 depicts an exemplary non-informative marker in a maternal only sample and a combination sample. Triangle: reference allele; circle: alternative allele.

Detailed Description of the Invention

[0018] Unless otherwise defined herein, scientific and technical terms used in this application shall have the meanings that are commonly understood by those of ordinary skill in the art. Generally, nomenclature used in connection with, and techniques of, cell and tissue culture, molecular biology, cell and cancer biology, neurobiology, neurochemistry, virology, immunology, microbiology, pharmacology, genetics and protein and nucleic acid chemistry, described herein, are those well known and commonly used in the art.

[0019] All of the above, and any other publications, patents and published patent applications referred to in this application are specifically incorporated by reference herein. In case of conflict, the present specification, including its specific definitions, will control.

[0020] Throughout this specification, the word "comprise" or variations such as "comprises" or "comprising" will be understood to imply the inclusion of a stated integer (or components) or group of integers (or components), but not the exclusion of any other integer (or components) or group of integers (or components).

[0021] The singular forms "a," "an," and "the" include the plurals unless the context clearly dictates otherwise.

[0022] The term "including" is used to mean "including but not limited to". "Including" and "including but not limited to" are used interchangeably.

[0023] A "patient", "subject", or "individual" are used interchangeably and refer to either a human or a non-human animal. These terms include mammals, such as humans, primates, livestock animals (e.g., bovines, porcines), companion animals (e.g., canines, felines) and rodents (e.g., mice and rats).

[0024] As used herein, a control subject refers to any individual that has not been diagnosed as having the disease or condition being assayed. The terms "normal control", "healthy control", and "not-diseased cells" likewise mean a sample (e.g., cells, serum, tissue) taken from a source (e.g., subject, control subject, cell line) that does not have the condition or disease being assayed and therefore may be used to determine the baseline for the condition or disorder being measured. It is also understood that the control subject, normal control, and healthy control, include data obtained and used as a standard, i.e. it can be used over and over again for multiple different subjects. In other words, for example, when comparing a subject sample to a control sample, the data from the control sample could have been obtained in a different set of experiments, for example, it could be an average obtained from a number of healthy subjects and not actually obtained at the time the data for the subject was obtained.

[0025] The term "diagnosis" as used herein refers to methods by which the skilled artisan can estimate and/or determine whether or not a patient is suffering from a given disease or condition. The skilled artisan often makes a diagnosis on the basis of one or more diagnostic indicators, e.g., a marker, the presence, absence, amount, or change in amount of which is indicative of the presence, severity, or absence of the condition.

[0026] The term "prognosis" as used herein refers to is used herein to refer to the likelihood of a disease or condition progression, including recurrence of a disease or condition.

[0027] The disclosure of the International Applications PCT/US09/31395, PCT/US11/45009, PCT/US11/44969, PCT/US11/44973, PCT/US11/44991, PCT/US11/45002, PCT/US11/44996, and PCT/US11/45018 are incorporated herein by reference for all purposes.

[0028] Each embodiment described herein may be combined with any other embodiment described herein.

Description of Methods of the Invention

[0029] By using a profile derived from a combination of sources, the skilled worker is able to capture data that is normally lost in the process of separating components out of a sample for testing. At the same time, the analytical components being combined are enriched for the markers used. Thus, the methods of the invention do not introduce unnecessary "noise" into the signal. Embodiments using a subject-specific profile comparison also eliminate the dependence on a population-derived average profile for a particular disease or condition. Using a population-derived average profile may introduce error into the detection or diagnosis of a particular disease or condition in the subject. Therefore, methods of this invention allow detection, diagnosis, and treatment to be personalized to the individual.

[0030] The methods of this invention have high specificity, sensitivity, and accuracy, and are capable of detecting disease or condition-specific markers present within a bodily fluid sample, cells or tissues. In some embodiments, the methods of this invention have improved specificity, sensitivity, and accuracy compared to current methods. The methods of the invention also reduce the problem of losing valuable signal when, for example, a patient sample is collected and separated into individual components, of which only one is chosen for testing. Accordingly, in certain aspects, the invention provides non-invasive assays for the early detection of a disease or condition, i.e., before the disease can be diagnosed by conventional diagnostic techniques, e.g., imaging techniques, and, therefore, provide a foundation for improved decision-making relative to the needs and strategies for intervention, prevention, and treatment of individuals with such disease or condition.

[0031] The present invention provides methods for diagnosing or aiding in the diagnosis of a disease or condition by comparing profiles (e.g., gene/protein/lipid/carbohydrate expression profiles, genotypes, gene copy number, gene dosage, DNA methylation, etc.) of disease or condition-associated markers (e.g., nucleic acids, proteins, lipids, carbohydrates, metabolites). One profile used for the comparison may be a profile from a sample comprising combinations of two or more different components (e.g., cell-free bodily fluids, phagocytic cells, circulating vesicles, and circulating diseased cells) from a subject. Alternatively, profile used for the comparison may be a profile from a sample comprising analytes isolated from,

e.g., cell-free bodily fluids, phagocytic cells, phagocytic cells having a DNA content more than 2n (>2n phagocytic cells), circulating vesicles, and circulating diseased cells from a subject. For example, a sample comprising analytes from phagocytes and a cell-free bodily fluid could be generated by isolating analytes (e.g., nucleic acids or proteins) from phagocytes (e.g., by lysing the cells and using affinity-based techniques to isolate the analytes) and analytes (e.g., nucleic acids or proteins) from a cell-free bodily fluid, and combining the analytes to create a sample useful in creating a profile. For ease of reference, a sample comprising combinations of two or more different components (e.g., cell-free bodily fluids, phagocytic cells, >2n phagocytic cells, circulating vesicles, and circulating diseased cells, or analytes isolated therefrom), may be referred to herein as a "combination sample." In some embodiments, the combination sample may comprise a lysate, or a fraction or portion of a lysate, of a population of phagocytic cells, a population of >2n phagocytic cells, a population of circulating vesicles, or a population of circulating diseased cells in place of the population of phagocytic cells, the population of >2n phagocytic cells, the population of circulating vesicles, or the population of circulating diseased cells, respectively. In some embodiments, the combination sample may comprise an analyte isolated from a lysate, or a fraction or portion of an analyte isolated from a lysate, of a population of phagocytic cells, a population of >2n phagocytic cells, a population of circulating vesicles, or a population of circulating diseased cells in place of the population of phagocytic cells, the population of >2n phagocytic cells, the population of circulating vesicles, or the population of circulating diseased cells, respectively. In some embodiments, the combination sample may comprise an analyte isolated from a cell-free bodily fluid, or a portion or fraction of an analyte isolated from a cell-free bodily fluid, rather than the cell-free bodily fluid itself. A control profile may be from phagocytic cells, phagocytic cells having a DNA content of 2n (=2n phagocytic cells), non-phagocytic cells, or control cells that are substantially free of cells affected by the disease or condition, taken from the same individual. Alternatively, the control profile may be a profile from a repository of markers of a disease or condition. In the context of this invention, "control cells that are substantially free of cells affected by the disease or condition" refers to a population of cells, as compared to circulating diseased cells, comprise significantly

fewer amounts of cells affected by the disease or condition. In some embodiments, "control cells that are substantially free of cells affected by the disease or condition" are cells that are at least 75%, 80%, 85%, 90%, 91%, 92%, 93%, 94%, 95%, 96%, 97%, 98%, 99%, 99.5%, or 100% free of cells affected by the disease or condition. In some embodiments, control cells that can be used in the methods of this invention are substantially free of fetal material (e.g., nucleic acids, proteins, and any analyte described herein), such as control cells that are at least 75%, 80%, 85%, 90%, 91%, 92%, 93%, 94%, 95%, 96%, 97%, 98%, 99%, 99.5%, or 100% free of fetal material.

[0032] This invention also provides methods for assessing the risk of developing a disease or condition, prognosing said disease, monitoring said disease progression or regression, assessing the efficacy of a treatment, or identifying a compound capable of ameliorating or treating said disease or condition.

[0033] In some embodiments, one or more components of whole blood that are substantially free of markers of a disease or condition may be isolated from a whole blood sample. In these embodiments, the one or more components of whole blood that are substantially free of markers may be used as a control sample (e.g., to determine a control profile), and the remaining portion of the whole blood sample may be used as an analytical sample (e.g., to determine an analytical profile). For example, a population of non-phagocytic cells may be isolated from a whole blood sample and used to determine a control profile, while the remaining portion of the whole blood sample is used to determine an analytical profile.

[0034] The methods of this invention can be used together with any known diagnostic methods, such as physical inspection, visual inspection, biopsy, scanning, histology, radiology, imaging, ultrasound, use of a commercial kit, genetic testing, immunological testing, analysis of bodily fluids, or monitoring neural activity.

[0035] Phagocytic cells that can be used in the methods of this invention include all types of cells that are capable of ingesting various types of substances (e.g., apoptotic cells, infectious agents, dead cells, viable cells, cell-free DNAs, cell-free RNAs, cell-free proteins). In some embodiments, the phagocytic cells are neutrophils, macrophages, monocytes, dendritic cells, foam cells, mast cells, eosinophils, or

keratinocytes. In some embodiments, the phagocytic cells can be a mixture of different types of phagocytic cells. In some embodiments, the phagocytic cells can be activated phagocytic cells., e.g., activated macrophages or neutrohils. In some embodiments, a phagocyte is a histiocyte, e.g., a Langerhans cell.

[0036] As used herein, ">2n phagocytic cells" refer to phagocytic cells that have a DNA content of greater than 2n, while "=2n phagocytic cells" refer to phagocytic cells that have a DNA content of 2n. According to the present invention, some phagocytic cells engulf live/dying/dead diseased cells (and sub-cellular fragments thereof) and/or cell-free disease-specific nucleic acids, proteins, carbohydrates and/or lipids present in bodily fluids. Such phagocytosis leads to the internalization of these disease markers into the phagocytic cell and, therefore, the DNA content of these phagocytic cells will become greater than 2n. By contrast, some phagocytic cells have not engulfed living/dying/dead diseased cells or fragments and/or cell-free disease-specific nucleic acids, proteins, lipids, and/or carbohydrates present in bodily fluids. The DNA contents of this group of phagocytic cells remain 2n. In some embodiments, the disease-specific markers (e.g., DNA with disease-specific mutations) can be expressed by the >2n phagocytic cells. For example, the mutated DNA of diseased cells is integrated into the normal DNA of the >2n phagocytic cells. The subsequent transcription of the "integrated" DNA of the >2n phagocytic cells into RNA and the translation of the latter into proteins produces a phenotype different from the phagocytic cells that have not phagocytosed the diseased cells (i.e., the =2n phagocytic cells). In other embodiments, the internalized disease-specific markers are not expressed by the >2n phagocytic cells. The markers may be translocated onto the membranes of the >2n phagocytic cells, or secreted out by the >2n phagocytic cells.

[0037] Circulating diseased cells that can be used in the methods of this invention include all types of circulating cells that may be affected by a disease or condition or infected by an infectious agent. A circulating cell refers to a cell present in the bodily fluid. A circulating cell may not necessarily circulate throughout the entire body or in the circulatory system. For example, a circulating cell may be present locally, such as in synovial fluid, or cerebrospinal fluid, or lymph fluid. A circulating diseased cell

may also be detached from a tissue or organ that has been affected by a disease or condition or infected by an infectious agent.

[0038] Cells that can be used as control cells in the methods of this invention includes all types of normal cells, or healthy cells, or cells that are substantially free of a disease or condition, or cells that are substantially free of an infectious agent. Control cells may be circulating cells or non-circulating cells (e.g., biopsied cells) that are representative of a normal or non-diseased state to which measurements on circulating diseased cells are compared to determine whether one or more diseased-associated marker is present in different levels between the circulating diseased cells and the control cells. The nature of the control cell may be a matter of design choice for a particular assay and may be derived or determined from normal tissue of the patient him- or herself.

[0039] In some embodiments, the circulating diseased cells are blood cells, tumor cells, lymphoma cells, fetal cells, apoptotic cells, epithelia cells, endothelial cells, stem cells, progenitor cells, mesenchymal cells, osteoblast cells, osteocytes, hematopoietic stem cells, foam cells, adipose cells, transcervical cells, circulating cardiocytes, circulating fibrocytes, circulating cancer stem cells, circulating myocytes, circulating cells from kidney, circulating cells from gastrointestinal tract, circulating cells from lung, circulating cells from reproductive organs, circulating cells from central nervous system, circulating hepatic cells, circulating cells from spleen, circulating cells from thymus, circulating cells from thyroid , circulating cells from an endocrine gland, circulating cells from parathyroid, circulating cells from pituitary, circulating cells from adrenal gland, circulating cells from islets of Langerhans, circulating cells from pancreas, circulating cells from hypothalamus, circulating cells from prostate tissues, circulating cells from breast tissues, circulating cells from circulating retinal cells, circulating ophthalmic cells, circulating auditory cells, circulating epidermal cells, or circulating cells from the urinary tract. In other embodiments, the circulating diseased cells can be a mixture of different types of circulating diseased cells.

[0040] The circulating diseased cells that can be used in the methods of this invention may be affected by various diseases or conditions. Exemplary diseases or conditions are a cardiovascular disease or condition, a kidney-associated disease or condition, a prenatal or pregnancy-related disease or condition, a neurological or neuropsychiatric disease or condition, an autoimmune or immune-related disease or condition, a cancer, an infectious disease or condition, a mitochondrial disorder, a respiratory-gastrointestinal tract disease or condition, a reproductive disease or condition, an ophthalmic disease or condition, a musculo-skeletal disease or condition, or a dermal disease or condition.

[0041] The circulating diseased cells that can be used in the methods of this invention may be infected by an infectious agent, such as a virus, bacteria, fungus, parasite, protozoan, infectious protein or microorganism.

[0042] In some embodiments, a cell useful in the methods of the invention (e.g., a circulating diseased cell, control cell not affected by the disease or condition, phagocytic cell, >2n phagocytic cell, =2n phagocytic cell, or non-phagocytic cell) is enucleated. Cells may be enucleated, for example, by using physical removal (e.g., via microneedle, optical tweezer, or aspiration), chemical treatments, photoablation, or ultraviolet irradiation.

[0043] As used herein, a "circulating vesicle" refers to a membrane-bound vesicle of cellular origin. The circulating vesicle may not necessarily circulate throughout the entire body or in the circulatory system. For example, the circulating vesicle may be present locally, such as in synovial fluid, or cerebrospinal fluid, or lymph fluid. In some embodiments, the circulating vesicles are selected from the group consisting of circulating microvesicles, apoptotic bodies, micro-particles, membrane-bound vesicles, multivesicular bodies, nanovesicles, microparticles, and ARRDC-1 mediated microvesicles (ARMM). In further embodiments, the circulating microvesicles are exosomes or urinary exosomes.

[0044] In some embodiments, a combination sample comprises two or more different components selected from the group consisting of: a cell-free bodily fluid isolated from a subject, a population of phagocytic cells isolated from the subject, a

population of >2n phagocytic cells isolated from the subject, a population of circulating vesicles isolated from the subject, and a population of circulating diseased cells isolated from the subject. In some embodiments, a combination sample comprises an analyte isolated from a cell-free bodily fluid isolated from the subject, an analyte isolated from a population of phagocytic cells isolated from the subject, an analyte isolated from a population of >2n phagocytic cells isolated from the subject, an analyte isolated from a population of circulating vesicles isolated from the subject, and an analyte isolated from a population of circulating diseased cells isolated from the subject.

[0045] As used herein, a "profile" of a marker of a disease or condition can broadly refer to any information concerning the marker. This information can be either qualitative (e.g., presence or absence) or quantitative (e.g., levels, copy numbers, or dosages). In some embodiments, a profile of a marker can indicate the absence of this marker. The profile can be a nucleic acid (e.g., DNA or RNA) profile, a protein profile, a lipid profile, a carbohydrate profile, a metabolite profile, or a combination thereof. A "marker" as used herein generally refers to an analyte which is differentially detectable in phagocytes and is indicative of the presence of a disease or condition. An analyte is differentially detectable if it can be distinguished quantitatively or qualitatively in phagocytes.

[0046] The methods of this invention can be applied to various diseases or conditions. Exemplary diseases or conditions are a cardiovascular disease or condition, a kidney-associated disease or condition, a prenatal or pregnancy-related disease or condition, a neurological or neuropsychiatric disease or condition, an autoimmune or immune-related disease or condition, a cancer, an infectious disease or condition, a pediatric disease, disorder, or condition, a mitochondrial disorder, a respiratory-gastrointestinal tract disease or condition, a reproductive disease or condition, an ophthalmic disease or condition, a musculo-skeletal disease or condition, or a dermal disease or condition.

[0047] As used herein, the term "cardiovascular disease or condition" refers to any condition that affects systems of heart or blood vessels (arteries and veins). Examples

of cardiovascular diseases include, but are not limited to myocardial infarction, coronary artery disease, percutaneous transluminal coronary angioplasty (PTCA), coronary artery bypass surgery (CABG), restenosis, peripheral arterial disease, stroke, abdominal aorta aneurysm, intracranial aneurysm, large artery atherosclerotic stroke, cardiogenic stroke, an early onset myocardial infarction, heart failure, pulmonary embolism, acute coronary syndrome (ACS), angina, cardiac hypertrophy, arteriosclerosis, myocarditis, pancarditis, endocarditis, hypertension, congestive heart failure, atherosclerosis, cerebrovascular disease, declining cardiac health, ischemic heart disease, pericarditis, cardiogenic shock, alcoholic cardiomyopathy, congenital heart disease, ischemic cardiomyopathy, hypertensive cardiomyopathy, valvular cardiomyopathy, inflammatory cardiomyopathy, cardiomyopathy secondary to a systemic metabolic disease, dilated cardiomyopathy, hypertrophic cardiomyopathy, arrhythmogenic right ventricular cardiomyopathy, restrictive cardiomyopathy, noncompaction cardiomyopathy, valvular heart disease, hypertensive heart disease, myocardial ischemic attack, unstable angina, myocardial rupture, cardiogenic shock, embolism, deep vein thrombosis, arrhythmia, arrhythmogenic right ventricular cardiomyopathy, diabetic cardiomyopathy, mitral regurgitation, mitral valve prolapse, peripheral vascular disease, artery disease, carotid artery disease, deep vein thrombosis, venous diseases, cerebrovascular disease, arterial aneurysm, left ventricular hypertrophy, hypertensive renal disease, hypertensive retinal disease, vasculitis, left main disease, arterial vascular disease, venous vascular disease, thrombosis of the microcirculation, a transient cerebrovascular accident, limb ischemia, aneurysm, thrombosis, superficial venous thrombosis, and deep venous thrombosis.

[0048] As used herein, the term "kidney-associated disease or condition" refers to any disease or condition that affects kidney or renal system. Examples of kidney-associated disease include, but are not limited to, chronic kidney diseases, primary kidney diseases, non-diabetic kidney diseases, glomerulonephritis, interstitial nephritis, diabetic kidney diseases, diabetic nephropathy, glomerulosclerosis, rapid progressive glomerulonephritis, renal fibrosis, Alport syndrome, IDDM nephritis, mesangial proliferative glomerulonephritis, membrano proliferative glomerulonephritis, crescentic glomerulonephritis, renal insterstitial fibrosis, focal

segmental glomerulosclerosis, membranous nephropathy, minimal change disease, pauci-immune rapid progressive glomerulonephritis, IgA nephropathy, polycystic kidney disease, Dent's disease, nephrocytinosis, Heymann nephritis, autosomal dominant (adult) polycystic kidney disease, autosomal recessive (childhood) polycystic kidney disease, acute kidney injury, nephrotic syndrome, renal ischemia, podocyte diseases or disorders, proteinuria, glomerular diseases, membranous glomerulonephritis, focal segmental glomerulonephritis, pre-eclampsia, eclampsia, kidney lesions, collagen vascular diseases, benign orthostatic (postural) proteinuria, IgM nephropathy, membranous nephropathy, sarcoidosis, diabetes mellitus, kidney damage due to drugs, Fabry's disease, aminoaciduria, Fanconi syndrome, hypertensive nephrosclerosis, interstitial nephritis, Sickle cell disease, hemoglobinuria, myoglobinuria, Wegener's Granulomatosis, Glycogen Storage Disease Type 1, chronic kidney disease, chronic renal failure, low Glomerular Filtration Rate (GFR), nephroangiosclerosis, lupus nephritis, ANCA-positive pauci-immune crescentic glomerulonephritis, chronic allograft nephropathy, nephrotoxicity, renal toxicity, kidney necrosis, kidney damage, glomerular and tubular injury, kidney dysfunction, nephritic syndrome, acute renal failure, chronic renal failure, proximal tubal dysfunction, acute kidney transplant rejection, chronic kidney transplant refection, non IgA mesangioproliferative glomerulonephritis, postinfectious glomerulonephritis, vasculitides with renal involvement of any kind, any hereditary renal disease, any interstitial nephritis, renal transplant failure, kidney cancer, kidney disease associated with other conditions (e.g., hypertension, diabetes, and autoimmune disease), Dent's disease, nephrocytinosis, Heymann nephritis, a primary kidney disease, a collapsing glomerulopathy, a dense deposit disease, a cryoglobulinemia-associated glomerulonephritis, an Henoch-Schónlein disease, a postinfectious glomerulonephritis, a bacterial endocarditis, a microscopic polyangitis, a Churg-Strauss syndrome, an anti-GBM-antibidy mediated glomerulonephritis, amyloidosis, a monoclonal immunoglobulin deposition disease, a fibrillary glomerulonephritis, an immunotactoid glomerulopathy, ischemic tubular injury, a medication-induced tubulo-interstitial nephritis, a toxic tubulo-interstitial nephritis, an infectious tubulo-interstitial nephritis, a bacterial pyelonephritis, a viral infectious tubulo-interstitial nephritis which results from a polyomavirus infection or an HIV infection, a

metabolic-induced tubulo-interstitial disease, a mixed connective disease, a cast nephropathy, a crystal nephropathy which may results from urate or oxalate or drug-induced crystal deposition, an acute cellular tubulo-interstitial allograft rejection, a tumoral infiltrative disease which results from a lymphoma or a post-transplant lymphoproliferative disease, an obstructive disease of the kidney, vascular disease, a thrombotic microangiopathy, a nephroangiosclerosis, an atheroembolic disease, a mixed connective tissue disease, a polyarteritis nodosa, a calcineurin-inhibitor induced- vascular disease, an acute cellular vascular allograft rejection, an acute humoral allograft rejection, early renal function decline (ERFD), end stage renal disease (ESRD), renal vein thrombosis, acute tubular necrosis, acute interstitial nephritis, established chronic kidney disease, renal artery stenosis, ischemic nephropathy, uremia, drug and toxin-induced chronic tubulointerstitial nephritis, reflux nephropathy, kidney stones, Goodpasture's syndrome, and hydronephrosis.

[0049] As used herein, the term "prenatal or pregnancy-related disease or condition" refers to any disease, disorder, or condition affecting a pregnant woman, embryo, or fetus. Prenatal or pregancy-related conditions can also refer to any disease, disorder, or condition that is associated with or arises, either directly or indirectly, as a result of pregnancy. These diseases or conditions can include any and all birth defects, congenital conditions, or hereditary diseases or conditions. Examples of prenatal or pregnancy-related diseases include, but are not limited to, Rhesus disease, hemolytic disease of the newborn, beta-thalassemia, sex determination, determination of pregnancy, a hereditary Mendelian genetic disorder, chromosomal aberrations, a fetal chromosomal aneuploidy, fetal chromosomal trisomy, fetal chromosomal monosomy, trisomy 8, trisomy 13 (Patau Syndrom), trisomy 16, trisomy 18 (Edwards syndrome), trisomy 21 (Down syndrome), X-chromosome linked disorders, trisomy X (XXX syndrome), monosomy X (Turner syndrome), XXY syndrome, XYY syndrome, XYY syndrome, XXXY syndrome, XXYY syndrome, XYYY syndrome, XXXXX syndrome, XXXXY syndrome, XXXYY syndrome, XXYYY syndrome, Fragile X Syndrome, fetal growth restriction, cystic fibrosis, a hemoglobinopathy, fetal death, fetal alcohol syndrome, sickle cell anemia, hemophilia, Klinefelter syndrome, dup(17)(p11.2p1.2) syndrome, endometriosis, Pelizaeus-Merzbacher disease, dup(22)(q11.2q11.2) syndrome, cat eye syndrome, cri-du-chat syndrome, Wolf-

Hirschhorn syndrome, Williams-Beuren syndrome, Charcot-Marie-Tooth disease, neuropathy with liability to pressure palsies, Smith-Magenis syndrome, neurofibromatosis, Alagille syndrome, Velocardiofacial syndrome, DiGeorge syndrome, steroid sulfatase deficiency, Prader-Willi syndrome, Kallmann syndrome, microphthalmia with linear skin defects, adrenal hypoplasia, glycerol kinase deficiency, Pelizaeus-Merzbacher disease, testis-determining factor on Y, azospermia (factor a), azospermia (factor b), azospermia (factor c), 1p36 deletion, phenylketonuria, Tay-Sachs disease, adrenal hyperplasia, Fanconi anemia, spinal muscular atrophy, Duchenne's muscular dystrophy, Huntington's disease, myotonic dystrophy, Robertsonian translocation, Angelman syndrome, tuberous sclerosis, ataxia telangieltasia, open spina bifida, neural tube defects, ventral wall defects, small-for-gestational-age, congenital cytomegalovirus, achondroplasia, Marfan's syndrome, congenital hypothyroidism, congenital toxoplasmosis, biotinidase deficiency, galactosemia, maple syrup urine disease, homocystinuria, medium-chain acyl Co-A dehydrogenase deficiency, structural birth defects, heart defects, abnormal limbs, club foot, anencephaly, arhinencephaly/holoprosencephaly, hydrocephaly, anophthalmos/microphthalmos, anotia/microtia, transposition of great vessels, tetralogy of Fallot, hypoplastic left heart syndrome, coarctation of aorta, cleft palate without cleft lip, cleft lip with or without cleft palate, oesophageal atresia/stenosis with or without fistula, small intestine atresia/stenosis, anorectal atresia/stenosis, hypospadias, indeterminate sex, renal agenesis, cystic kidney, preaxial polydactyly, limb reduction defects, diaphragmatic hernia, blindness, cataracts, visual problems, hearing loss, deafness, X-linked adrenoleukodystrophy, Rett syndrome, lysosomal disorders, cerebral palsy, autism, aglossia, albinism, ocular albinism, oculocutaneous albinism, gestational diabetes, Arnold-Chiari malformation, CHARGE syndrome, congenital diaphragmatic hernia, brachydactlia, aniridia, cleft foot and hand, heterochromia, Dwarnian ear, Ehlers Danlos syndrome, epidermolysis bullosa, Gorham's disease, Hashimoto's syndrome, hydrops fetalis, hypotonia, Klippel-Feil syndrome, muscular dystrophy, osteogenesis imperfecta, progeria, Smith Lemli Opitz symdrom, chromatelopsia, X-linked lymphoproliferative disease, omphalocele, gastroschisis, pre-eclampsia, eclampsia, pre-term labor, premature birth, miscarriage,

delayed intrauterine growth, ectopic pregnancy, hyperemesis gravidarum, morning sickness, or likelihood for successful induction of labor.

[0050] As used herein, the term "a neurological or neuropsychiatric disease or condition" refers to any disease or condition that affects nervous systems. Examples of neurological or neuropsychiatric diseases or conditions include, but are not limited to, head trauma, stroke, stroke, ischemic stroke, hemorrhagic stroke, subarachnoid hemorrhage, intra cranial hemorrhage, transient ischemic attack, vascular dementia, corticobasal ganglionic degeneration, encephalitis, epilepsy, Landau-Kleffner syndrome, hydrocephalus, pseudotumor cerebri, thalamic diseases, meningitis, myelitis, movement disorders, essential tremor, spinal cord diseases, syringomyelia, Alzheimer's disease (early onset), Alzheimer's disease (late onset), multi-infarct dementia, Pick's disease, Huntingdon's disease, Parkinson's disease, Parkinson syndromes, dementia, frontotemporal dementia, corticobasal degeneration, multiple system atrophy, progressive supranuclear palsy, Lewy body disease, Creutzfeldt-Jakob disease, Dandy-Walker syndrome, Friedreich ataxia, Machado-Joseph disease, migraine, schizophrenia, mood disorders and depression. dementia with lewy bodies (DLB), frontotemporal dementia (FTD), various forms of vascular dementia (VD), subcortical vascular dementia (Binswanger's disease), autism, developmental retardations, motor neuron diseases, amyotrophic lateral sclerosis (ALS), neuronal or brain damage, hypoxia of the brain, cerebral palsy (CP), memory disorders, movement disorders, corticalbasal ganglionic degeneration, forms of multiple system atrophy, stroke-related disorders, cerebrovascular accidents, post-irradiation encephalopathy with seizures, vascular Parkinsonism, thalamic cerebrovascular accidents, chronic inflammatory demyelinating polyneuropathy, alcohol related dementia, semantic dementia, ataxia, atypical Parkinsonism, dystonia, progressive supranuclear palsy, essential tremor, mild cognitive impairment, amyotrophic lateral sclerosis, multiple sclerosis, neuropathies, Pick's disease, congophilic amyloid angiopathy, Creutzfeldt-Jakob Disease, AIDS dementia complex, depression, anxiety disorder, phobia, Bell's Palsy, epilepsy, encephalitis, neuromuscular disorders, neurooncological disorders, brain tumors, neurovascular disorders, neuroimmunological disorders, neurootological disease, neurotrauma including spinal cord injury, pain including neuropathic pain, pediatric neurological and

neuropsychiatric disorders, sleep disorders, Tourette syndrome, corticalbasal ganglionic degeneration, alcohol related dementia, semantic dementia, Alzheimer's disease combined with multi-infarct dementia, Alzheimer's disease combined with Lewy body dementia, Parkinson's disease combined with Lewy body dementia, Alzheimer's and Parkinson's disease combined with Lewy body dementia, frontotemporal dementia combined with chronic inflammatory demyelinating polyneuropathy, attention deficit hyperactivity disorder, schizophrenia, obsessive-compulsive disorder, mental retardation, autistic spectrum disorders, opsoclonus-myoclonus syndrome (OMS) seizures, articulation disorder, learning disabilities (i.e., reading or arithmetic), verbal or performance aptitude deficits, attention deficit disorder, amyloid diseases, prion diseases, Tauopathies, Alpha-Synucleinopathies, addictive states such as those caused by at least one of: cocaine, nicotine, alcohol, food, ecstasy, kat, caffeine, opium, heroin, marijuana, amphetamine, methamphetamine or gambling, and Fabry's disease.

[0051] As used herein, the term "an autoimmune or immune-related disease or condition" refers to any disease or condition that affects the function of immune systems. Examples of autoimmune or immune-related diseases or conditions include, but are not limited to, antiphospholipid syndrome, systemic lupus erythematosus, rheumatoid arthritis, autoimmune vasculitis, celiac disease, autoimmune thyroiditis, post-transfusion immunization, maternal-fetal incompatibility, transfusion reactions, immunological deficiency such IgA deficiency, common variable immunodeficiency, drug-induced lupus, diabetes mellitus, Type I diabetes, Type II diabetes, juvenile onset diabetes, juvenile rheumatoid arthritis, psoriatic arthritis, multiple sclerosis, immunodeficiency, allergies, asthma, psoriasis, atopic dermatitis, allergic contact dermatitis, chronic skin diseases, amyotrophic lateral sclerosis, chemotherapy-induced injury, graft-vs-host diseases, bone marrow transplant rejection, Ankylosing spondylitis, atopic eczema, Pemphigus, Behcet's disease, chronic fatigue syndrome fibromyalgia, chemotherapy-induced injury, myasthenia gravis, glomerulonephritis, allergic retinitis, systemic sclerosis, subacute cutaneous lupus erythematosus, cutaneous lupus erythematosus including chilblain lupus erythematosus, Sjogren's syndrome, autoimmune nephritis, autoimmune vasculitis, autoimmune hepatitis, autoimmune carditis, autoimmune encephalitis, autoimmune mediated hematological

diseases, lc-SSc (limited cutaneous form of scleroderma), dc-SSc (diffused cutaneous form of scleroderma), autoimmune thyroiditis (AT), Grave's disease (GD), myasthenia gravis, multiple sclerosis (MS), ankylosing spondylitis. transplant rejection, immune aging, rheumatic/autoimmune diseases, mixed connective tissue disease, spondyloarthropathy, psoriasis, psoriatic arthritis, myositis, scleroderma, dermatomyositis, autoimmune vasculitis, mixed connective tissue disease, idiopathic thrombocytopenic purpura, Crohn's disease, human adjuvant disease, osteoarthritis, juvenile chronic arthritis, a spondyloarthropathy, an idiopathic inflammatory myopathy, systemic vasculitis, sarcoidosis, autoimmune hemolytic anemia, autoimmune thrombocytopenia, thyroiditis, immune-mediated renal disease, a demyelinating disease of the central or peripheral nervous system, idiopathic demyelinating polyneuropathy, Guillain-Barre syndrome, a chronic inflammatory demyelinating polyneuropathy, a hepatobiliary disease, infectious or autoimmune chronic active hepatitis, primary biliary cirrhosis, granulomatous hepatitis, sclerosing cholangitis, inflammatory bowel disease, gluten-sensitive enteropathy, Whipple's disease, an autoimmune or immune-mediated skin disease, a bullous skin disease, erythema multiforme, allergic rhinitis, atopic dermatitis, food hypersensitivity, urticaria, an immunologic disease of the lung, eosinophilic pneumonias, idiopathic pulmonary fibrosis, hypersensitivity pneumonitis, a transplantation associated disease, graft rejection or graft-versus-host-disease, psoriatic arthritis, psoriasis, dermatitis, polymyositis/dermatomyositis, toxic epidermal necrolysis, systemic scleroderma and sclerosis, responses associated with inflammatory bowel disease, Crohn's disease, ulcerative colitis, respiratory distress syndrome, adult respiratory distress syndrome (ARDS), meningitis, encephalitis, uveitis, colitis, glomerulonephritis, allergic conditions, eczema, asthma, conditions involving infiltration of T cells and chronic inflammatory responses, atherosclerosis, autoimmune myocarditis, leukocyte adhesion deficiency, allergic encephalomyelitis, immune responses associated with acute and delayed hypersensitivity mediated by cytokines and T-lymphocytes, tuberculosis, sarcoidosis, granulomatosis including Wegener's granulomatosis, agranulocytosis, vasculitis (including ANCA), aplastic anemia, Diamond Blackfan anemia, immune hemolytic anemia including autoimmune hemolytic anemia (AIHA), pernicious anemia, pure red cell aplasia (PRCA), Factor VIII deficiency, hemophilia

A, autoimmune neutropenia, pancytopenia, leukopenia, diseases involving leukocyte diapedesis, central nervous system (CNS) inflammatory disorders, multiple organ injury syndrome, mysathenia gravis, antigen-antibody complex mediated diseases, anti-glomerular basement membrane disease, anti-phospholipid antibody syndrome, allergic neuritis, Bechet disease, Castleman's syndrome, Goodpasture's syndrome, Lambert-Eaton Myasthenic Syndrome, Reynaud's syndrome, Sjorgen's syndrome, Stevens-Johnson syndrome, pemphigoid bullous, pemphigus, autoimmune polyendocrinopathies, Reiter's disease, stiff-man syndrome, giant cell arteritis, immune complex nephritis, IgA nephropathy, IgM polyneuropathies or IgM mediated neuropathy, idiopathic thrombocytopenic purpura (ITP), thrombotic throbocytopenic purpura (TTP), autoimmune thrombocytopenia, autoimmune disease of the testis and ovary including autoimmune orchitis and oophoritis, primary hypothyroidism, autoimmune endocrine diseases including autoimmune thyroiditis, chronic thyroiditis (Hashimoto's Thyroiditis), subacute thyroiditis, idiopathic hypothyroidism, Addison's disease, Grave's disease, autoimmune polyglandular syndromes (or polyglandular endocrinopathy syndromes), Sheehan's syndrome, autoimmune hepatitis, lymphoid interstitial pneumonitis (HIV), bronchiolitis obliterans (non-transplant) vs NSIP, Guillain-Barre' Syndrome, large vessel vasculitis (including polymyalgia rheumatica and giant cell (Takayasu's) arteritis), medium vessel vasculitis (including Kawasaki's disease and polyarteritis nodosa), ankylosing spondylitis, Berger's disease (IgA nephropathy), rapidly progressive glomerulonephritis, primary biliary cirrhosis, Celiac sprue (gluten enteropathy), cryoglobulinemia, and amyotrophic lateral sclerosis (ALS).

[0052] As used herein, the term "cancer" refers to various types of malignant neoplasms, most of which can invade surrounding tissues, and may metastasize to different sites (see, for example, PDR Medical Dictionary, 1st edition (1995), incorporated herein by reference in its entirety for all purposes). The terms "neoplasm" and "tumor" refer to an abnormal tissue that grows by cellular proliferation more rapidly than normal and continues to grow after the stimuli that initiated proliferation is removed. Such abnormal tissue shows partial or complete lack of structural organization and functional coordination with the normal tissue which may be either benign (i.e., benign tumor) or malignant (i.e., malignant tumor).

Examples of general categories of cancer include, but are not limited to, carcinomas (i.e., malignant tumors derived from epithelial cells such as, for example, common forms of breast, prostate, lung and colon cancer), sarcomas (i.e., malignant tumors derived from connective tissue or mesenchymal cells), lymphomas (i.e., malignancies derived from hematopoietic cells), leukemias (i.e., malignancies derived from hematopoietic cells), germ cell tumors (i.e., tumors derived from totipotent cells. In adults most often found in the testicle or ovary; in fetuses, babies and young children, most often found on the body midline, particularly at the tip of the tailbone), blastic tumors (i.e., a typically malignant tumor which resembles an immature or embryonic tissue) and the like. Examples of the types of neoplasms intended to be encompassed by the present invention include but are not limited to those neoplasms associated with cancers of neural tissue, blood forming tissue, breast, skin, bone, prostate, ovaries, uterus, cervix, liver, lung, brain, larynx, gallbladder, pancreas, rectum, parathyroid, thyroid, adrenal gland, immune system, head and neck, colon, stomach, bronchi, and/or kidneys.

[0053] As used herein, the term "an infectious disease or condition" refers to any disease or condition that results from an infectious agent. Infectious agents include, but are not limited to bacteria, viruses, fungi, protozoa, infectious proteins, parasitic microbes, and other parasites. Examples of infectious diseases or conditions include, but are not limited to, bacterial infections, viral infections, fungal infections, protozoan infections, parasitic infections, hepatitis (e.g., hepatitis A, B, C, D, and E), herpes, influenza, human papillomavirus (HPV) infection, AIDS, anthrax, pneumonia (bacterial or viral), cellulitis, human parainfluenza, the common cold, Legionnaires' disease (Legionellosis), cholera, Creutzfeldt-Jakob disease (CJD), variant Creutzfeldt-Jakob disease (vCJD), fatal familial insomnia (FFI), Gerstmann-Sträussler-Scheinker (GSS) syndrome, Chlamydia, chicken pox, ebola hemorrhagic fever, Dengue fever, giardiasis, Lyme disease, malaria, measles, mumps, rubella, pertussis, gonorrhea, staphylococcal infection, streptococcal infection, pneumococcal infection, rabies, helicobacter pylori infection, respiratory syncitial virus infection, Rocky Mountain spotted fever, SARS, sepsis, tuberculosis, and West Nile fever.

[0054] Viruses include, but are not limited to, DNA or RNA animal viruses. As used herein, RNA viruses include, but are not limited to, virus families such as Picornaviridae (e.g., polioviruses), Reoviridae (e.g., rotaviruses), Togaviridae (e.g., encephalitis viruses, yellow fever virus, rubella virus), Orthomyxoviridae (e.g., influenza viruses), Paramyxoviridae (e.g., respiratory syncytial virus, measles virus, mumps virus, parainfluenza virus), Rhabdoviridae (e.g., rabies virus), Coronaviridae, Bunyaviridae, Flaviviridae, Filoviridae, Arenaviridae, Bunyaviridae and Retroviridae (e.g., human T cell lymphotropic viruses (HTLV), human immunodeficiency viruses (HIV)). As used herein, DNA viruses include, but are not limited to, virus families such as Papovaviridae (e.g., papilloma viruses), Adenoviridae (e.g., adenovirus), Herpesviridae (e.g., herpes simplex viruses), and Poxviridae (e.g., variola viruses).

[0055] Bacteria include, but are not limited to, gram positive bacteria, gram negative bacteria, acid-fast bacteria and the like. Gram positive bacteria include, but are not limited to, Actinomedurae, Actinomyces israelii, Bacillus anthracis, Bacillus cereus, Clostridium botulinum, Clostridium difficile, Clostridium perfringens, Clostridium tetani, Corynebacterium, Enterococcus faecalis, Listeria monocytogenes, Nocardia, Propionibacterium acnes, Staphylococcus aureus, Staphylococcus epiderm, Streptococcus mutans, Streptococcus pneumoniae and the like. Gram negative bacteria include, but are not limited to, Afipia felis, Bacteriodes, Bartonella bacilliformis, Bortadella pertussis, Borrelia burgdorferi, Borrelia recurrentis, Brucella, Calymmatobacterium granulomatis, Campylobacter, Escherichia coli, Francisella tularensis, Gardnerella vaginalis, Haemophilius aegyptius, Haemophilius ducreyi, Haemophilius influenziae, Heliobacter pylori, Legionella pneumophila, Leptospira interrogans, Neisseria meningitidia, Porphyromonas gingivalis, Providencia sturti, Pseudomonas aeruginosa, Salmonella enteridis, Salmonella typhi, Serratia marcescens, Shigella boydii, Streptobacillus moniliformis, Streptococcus pyogenes, Treponema pallidum, Vibrio cholerae, Yersinia enterocolitica, Yersinia pestis and the like. As used herein, acid-fast bacteria include, but are not limited to, Myobacterium avium, Myobacterium leprae, Myobacterium tuberculosis and the like. Other bacteria not falling into the other three categories include, but are not limited to, Bartonella henseiae, Chlamydia psittaci, Chlamydia trachomatis, Coxiella burnetii, Mycoplasma pneumoniae, Rickettsia akari, Rickettsia prowazekii, Rickettsia rickettsii, Rickettsia

tsutsugamushi, Rickettsia typhi, Ureaplasma urealyticum, Diplococcus pneumoniae, Ehrlichia chafensis, Enterococcus faecium, Meningococci and the like.

[0056] As used herein, fungi include, but are not limited to, Aspergilli, Candidae, Candida albicans, Coccidioides immitis, Cryptococci, and combinations thereof.

[0057] As used herein, parasitic microbes include, but are not limited to, Balantidium coli, Cryptosporidium parvum, Cyclospora cayatanensis, Encephalitozoa, Entamoeba histolytica, Enterocytozoon bieneusi, Giardia lamblia, Leishmaniae, Plasmodii, Toxoplasma gondii, Trypanosomae, trapezoidal amoeba and the like. Other parasites include worms (e.g., helminthes), particularly parasitic worms including, but not limited to, Nematoda (roundworms, e.g., whipworms, hookworms, pinworms, ascarids, filarids and the like), and Cestoda (e.g., tapeworms).

[0058] As used herein, infectious proteins include prions (e.g., PrP^{Sc} forms, the CJD prion, the vCJD prion, the FFI prion, and the GSS prion).

[0059] As used herein, the term "a pediatric disease, disorder, or condition" refers to any disease or disorder that affects infants or children or that begins during development or childhood. Examples of pediatric diseases, disorders, and conditions include, but are not limited to, autism, Kawasaki's disease, congenital deafness, pediatric cancers, Type I diabetes, congenital heart defects, tetralogy of Fallot, Duchenne Muscular Dystrophy, osteogenesis importfect, Krabe disease, Pompe disease, Gaucher disease, Fabry disease, Wolff-Parkinson-White syndrome, Hirschsprung's disease, Crohn's disease, Eagle-Barrett Syndrome, cystic fibrosis, irritable bowel syndrome, and cerebral palsy. Examples of pediatric conditions also include genetic attributes of the developing fetus. For example, a pediatric conditions include, but are not limited to, intelligence, eye color, hair color, and muscle type.

[0060] As used herein, "treating" a disease or condition refers to taking steps to obtain beneficial or desired results, including clinical results. Beneficial or desired clinical results include, but are not limited to, alleviation or amelioration of one or more symptoms associated with diseases or conditions.