Electronic Acknowledgement Receipt

EFS ID:	20691707
Application Number:	14540917
International Application Number:	
Confirmation Number:	7076
Title of Invention:	COMPOSITIONS AND METHODS FOR IDENTIFICATION OF A DUPLICATE SEQUENCING READ
First Named Inventor/Applicant Name:	Douglas AMORESE
Customer Number:	58249
Filer:	Angie S.Y. Mah
Filer Authorized By:	
Attorney Docket Number:	NUGN-001/01US
Receipt Date:	13-NOV-2014
Filing Date:	
Time Stamp:	18:19:01
Application Type:	Utility under 35 USC 111(a)

Payment information:

Submitted with Payment	yes
Payment Type	Deposit Account
Payment was successfully received in RAM	$5820
RAM confirmation Number	4934
Deposit Account	501283
Authorized User	

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

Charge any Additional Fees required under 37 C.F.R. Section 1.16 (National application filing, search, and examination fees)

Charge any Additional Fees required under 37 C.F.R. Section 1.17 (Patent application and reexamination processing fees)

Charge any Additional Fees required under 37 C.F.R. Section 1.19 (Document supply fees)

Charge any Additional Fees required under 37 C.F.R. Section 1.20 (Post Issuance fees)

Charge any Additional Fees required under 37 C.F.R. Section 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	Transmittal of New Application	NUGN-001_01US_UtilityTransmittal.pdf	172100 3346403fd2f14f20763898f70c724a1778f746db	no	1
Warnings:					
Information:					
2	TrackOne Request	NUGN-001_01US_TrackOneTransmittal.pdf	135963 91825e519d517729ef1f6d96a2504778db586b4a	no	1
Warnings:					
Information:					
3	Application Data Sheet	NUGN-001_01US_ADS.pdf	111444 c7f45f621f9dd878c7ce4f71890e4e52371fca9a	no	6
Warnings:					
Information:					
This is not an USPTO supplied ADS fillable form					
4		NUGN-001_01US_Specifcation.pdf	254432 5c13358ecaa123e6469567a44e06e8d48c925771	yes	44

Multipart Description/PDF files in .zip description

Document Description	Start	End
Specification	1	40
Claims	41	43
Abstract	44	44

Warnings:					
Information:					
5	Drawings-other than black and white line drawings	NUGN-001_01US_Drawings.pdf	116178 40d64004a908d38384740853fb9d29e73a77b4fc	no	6
Warnings:					
Information:					
6	Sequence Listing (Text File)	NUGN_001_01US_SeqList_ST25.txt	2809	no	0

Warnings:					
Information:					
7	Fee Worksheet (SB06)	fee-info.pdf	41712 7927943e0c25baae0ae7bdbef708110e664e020e	no	2
Warnings:					
Information:					
			Total Files Size (in bytes):	834638	

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.

PTO/AIA/14 (12-13)
Approved for use through 01/31/2014. 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 contains a valid OMB control number.

Application Data Sheet 37 CFR 1.76	Attorney Docket Number	NUGN-001/01US
	Application Number	
Title of Invention	COMPOSITIONS AND METHODS FOR IDENTIFICATION OF A DUPLICATE SEQUENCING READ	

The application data sheet is part of the provisional or nonprovisional application for which it is being submitted. The following form contains the bibliographic data arranged in a format specified by the United States Patent and Trademark Office as outlined in 37 CFR 1.76.
This document may be completed electronically and submitted to the Office in electronic format using the Electronic Filing System (EFS) or the document may be printed and included in a paper filed application.

Secrecy Order 37 CFR 5.2

☐ Portions or all of the application associated with this Application Data Sheet may fall under a Secrecy Order pursuant to 37 CFR 5.2 (Paper filers only. Applications that fall under Secrecy Order may not be filed electronically.)

Inventor Information:

Inventor 1 Remove

Legal Name

Prefix	Given Name	Middle Name	Family Name	Suffix
	Douglas		AMORESE	

Residence Information (Select One) ◉ US Residency ○ Non US Residency ○ Active US Military Service

City	Los Altos	State/Province	CA	Country of Residence	US

Mailing Address of Inventor:

Address 1	357 South El Monte Avenue		
Address 2			
City	Los Altos	State/Province	CA
Postal Code		Country	US

Inventor 2 Remove

Legal Name

Prefix	Given Name	Middle Name	Family Name	Suffix
	Jonathan		SCOLNICK	

Residence Information (Select One) ◉ US Residency ○ Non US Residency ○ Active US Military Service

City	San Francisco	State/Province	CA	Country of Residence	US

Mailing Address of Inventor:

Address 1	451 Kansas Street, #518		
Address 2			
City	San Francisco	State/Province	CA
Postal Code		Country	US

Inventor 3 Remove

Legal Name

Prefix	Given Name	Middle Name	Family Name	Suffix
	Ben		SCHROEDER	

Residence Information (Select One) ◉ US Residency ○ Non US Residency ○ Active US Military Service

IDT Exhibit 1018.0004

PTO/AIA/14 (03-13)
Approved for use through 01/31/2014. 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 contains a valid OMB control number.

Application Data Sheet 37 CFR 1.76	Attorney Docket Number	NUGN-001/01US
	Application Number	
Title of Invention	COMPOSITIONS AND METHODS FOR IDENTIFICATION OF A DUPLICATE SEQUENCING READ	

City	San Mateo	State/Province	CA	Country of Residence	US

Mailing Address of Inventor:

Address 1	1964 Shoreview Avenue		
Address 2			
City	San Mateo	State/Province	CA
Postal Code		Country	US

All Inventors Must Be Listed - Additional Inventor Information blocks may be generated within this form by selecting the **Add** button. Add

Correspondence Information:

Enter either Customer Number or complete the Correspondence Information section below.
For further information see 37 CFR 1.33(a).

☐ **An Address is being provided for the correspondence Information of this application.**

Customer Number	58249		
Email Address	zpatdcdocketing@cooley.com	Add Email	Remove Email

Application Information:

Title of the Invention	COMPOSITIONS AND METHODS FOR IDENTIFICATION OF A DUPLICATE SEQUENCING READ		
Attorney Docket Number	NUGN-001/01US	Small Entity Status Claimed	☐
Application Type	Nonprovisional		
Subject Matter	Utility		
Total Number of Drawing Sheets (if any)	6	Suggested Figure for Publication (if any)	

Filing By Reference :

Only compete this section when filing an application by reference under 35 U.S.C. 111(c) and 37 CFR 1.57(a). Do not complete this section if application papers including a specification and any drawings are being filed. Any domestic benefit or foreign priority information must be provided in the appropriate section(s) below (i.e., "Domestic Benefit/National Stage Information" and "Foreign Priority Information").

For the purposes of a filing date under 37 CFR 1.53(b), the description and any drawings of the present application are replaced by this reference to the previously filed application, subject to conditions and requirements of 37 CFR 1.57(a).

Application number of the previously filed application	Filing date (YYYY-MM-DD)	Intellectual Property Authority or Country

Publication Information:

☐ Request Early Publication (Fee required at time of Request 37 CFR 1.219)

☐ **Request Not to Publish.** I hereby request that the attached application not be published under 35 U.S.C. 122(b) and certify that the invention disclosed in the attached application **has not and will not** be the subject of an application filed in another country, or under a multilateral international agreement, that requires publication at eighteen months after filing.

PTO/AIA/14 (03-13)
Approved for use through 01/31/2014. 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 contains a valid OMB control number.

Application Data Sheet 37 CFR 1.76	Attorney Docket Number	NUGN-001/01US
	Application Number	
Title of Invention	COMPOSITIONS AND METHODS FOR IDENTIFICATION OF A DUPLICATE SEQUENCING READ	

Representative Information:

Representative information should be provided for all practitioners having a power of attorney in the application. Providing this information in the Application Data Sheet does not constitute a power of attorney in the application (see 37 CFR 1.32).
Either enter Customer Number or complete the Representative Name section below. If both sections are completed the customer Number will be used for the Representative Information during processing.

Please Select One:	(●) Customer Number	( ) US Patent Practitioner	( ) Limited Recognition (37 CFR 11.9)
Customer Number	58249		

Domestic Benefit/National Stage Information:

This section allows for the applicant to either claim benefit under 35 U.S.C. 119(e), 120, 121, or 365(c) or indicate National Stage entry from a PCT application. Providing this information in the application data sheet constitutes the specific reference required by 35 U.S.C. 119(e) or 120, and 37 CFR 1.78.
When referring to the current application, please leave the application number blank.

Prior Application Status	Pending		Remove
Application Number	Continuity Type	Prior Application Number	Filing Date (YYYY-MM-DD)
	Claims benefit of provisional	61903826	2013-11-13

Additional Domestic Benefit/National Stage Data may be generated within this form by selecting the **Add** button.

Foreign Priority Information:

This section allows for the applicant to claim priority to a foreign application. Providing this information in the application data sheet constitutes the claim for priority as required by 35 U.S.C. 119(b) and 37 CFR 1.55(d). When priority is claimed to a foreign application that is eligible for retrieval under the priority document exchange program (PDX)[i] the information will be used by the Office to automatically attempt retrieval pursuant to 37 CFR 1.55(h)(1) and (2). Under the PDX program, applicant bears the ultimate responsibility for ensuring that a copy of the foreign application is received by the Office from the participating foreign intellectual property office, or a certified copy of the foreign priority application is filed, within the time period specified in 37 CFR 1.55(g)(1).

			Remove
Application Number	Country[i]	Filing Date (YYYY-MM-DD)	Access Code[i] (if applicable)

Additional Foreign Priority Data may be generated within this form by selecting the **Add** button.

IDT Exhibit 1018.0006

PTO/AIA/14 (03-13)
Approved for use through 01/31/2014. 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 contains a valid OMB control number.

Application Data Sheet 37 CFR 1.76	Attorney Docket Number	NUGN-001/01US
	Application Number	
Title of Invention	COMPOSITIONS AND METHODS FOR IDENTIFICATION OF A DUPLICATE SEQUENCING READ	

Statement under 37 CFR 1.55 or 1.78 for AIA (First Inventor to File) Transition Applications

☐ This application (1) claims priority to or the benefit of an application filed before March 16, 2013 and (2) also contains, or contained at any time, a claim to a claimed invention that has an effective filing date on or after March 16, 2013.
NOTE: By providing this statement under 37 CFR 1.55 or 1.78, this application, with a filing date on or after March 16, 2013, will be examined under the first inventor to file provisions of the AIA.

Authorization to Permit Access:

☒ Authorization to Permit Access to the Instant Application by the Participating Offices

If checked, the undersigned hereby grants the USPTO authority to provide the European Patent Office (EPO), the Japan Patent Office (JPO), the Korean Intellectual Property Office (KIPO), the World Intellectual Property Office (WIPO), and any other intellectual property offices in which a foreign application claiming priority to the instant patent application is filed access to the instant patent application. See 37 CFR 1.14(c) and (h). This box should not be checked if the applicant does not wish the EPO, JPO, KIPO, WIPO, or other intellectual property office in which a foreign application claiming priority to the instant patent application is filed to have access to the instant patent application.

In accordance with 37 CFR 1.14(h)(3), access will be provided to a copy of the instant patent application with respect to: 1) the instant patent application-as-filed; 2) any foreign application to which the instant patent application claims priority under 35 U.S.C. 119(a)-(d) if a copy of the foreign application that satisfies the certified copy requirement of 37 CFR 1.55 has been filed in the instant patent application; and 3) any U.S. application-as-filed from which benefit is sought in the instant patent application.

In accordance with 37 CFR 1.14(c), access may be provided to information concerning the date o f filing this Authorization.

Applicant Information:

Providing assignment information in this section does not substitute for compliance with any requirement of part 3 of Title 37 of CFR to have an assignment recorded by the Office.

PTO/AIA/14 (03-13)
Approved for use through 01/31/2014. 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 contains a valid OMB control number.

Application Data Sheet 37 CFR 1.76	Attorney Docket Number	NUGN-001/01US
	Application Number	
Title of Invention	COMPOSITIONS AND METHODS FOR IDENTIFICATION OF A DUPLICATE SEQUENCING READ	

Applicant 1

If the applicant is the inventor (or the remaining joint inventor or inventors under 37 CFR 1.45), this section should not be completed. The information to be provided in this section is the name and address of the legal representative who is the applicant under 37 CFR 1.43; or the name and address of the assignee, person to whom the inventor is under an obligation to assign the invention, or person who otherwise shows sufficient proprietary interest in the matter who is the applicant under 37 CFR 1.46. If the applicant is an applicant under 37 CFR 1.46 (assignee, person to whom the inventor is obligated to assign, or person who otherwise shows sufficient proprietary interest) together with one or more joint inventors, then the joint inventor or inventors who are also the applicant should be identified in this section.

Clear

○ Assignee	○ Legal Representative under 35 U.S.C. 117	○ Joint Inventor
◉ Person to whom the inventor is obligated to assign.	○ Person who shows sufficient proprietary interest	

If applicant is the legal representative, indicate the authority to file the patent application, the inventor is:

Name of the Deceased or Legally Incapacitated Inventor :

If the Applicant is an Organization check here. ☒

Organization Name	NUGEN TECHNOLOGIES, INC.		
Mailing Address Information For Applicant:			
Address 1	201 Industrial Road		
Address 2	Suite 310		
City	San Carlos	**State/Province**	CA
Country	US	Postal Code	94070
Phone Number		Fax Number	
Email Address			

Additional Applicant Data may be generated within this form by selecting the Add button.

Assignee Information including Non-Applicant Assignee Information:

Providing assignment information in this section does not subsitute for compliance with any requirement of part 3 of Title 37 of CFR to have an assignment recorded by the Office.

Assignee 1

Complete this section if assignee information, including non-applicant assignee information, is desired to be included on the patent application publication . An assignee-applicant identified in the "Applicant Information" section will appear on the patent application publication as an applicant. For an assignee-applicant, complete this section only if identification as an assignee is also desired on the patent application publication.

If the Assignee or Non-Applicant Assignee is an Organization check here. ☐

PTO/AIA/14 (03-13)
Approved for use through 01/31/2014. 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 contains a valid OMB control number.

Application Data Sheet 37 CFR 1.76	Attorney Docket Number	NUGN-001/01US
	Application Number	
Title of Invention	COMPOSITIONS AND METHODS FOR IDENTIFICATION OF A DUPLICATE SEQUENCING READ	

Prefix	**Given Name**	Middle Name	**Family Name**	Suffix

Mailing Address Information For Assignee including Non-Applicant Assignee:

Address 1			
Address 2			
City		**State/Province**	
Country i		Postal Code	
Phone Number		Fax Number	
Email Address			

Additional Assignee or Non-Applicant Assignee Data may be generated within this form by selecting the Add button.

Signature:

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

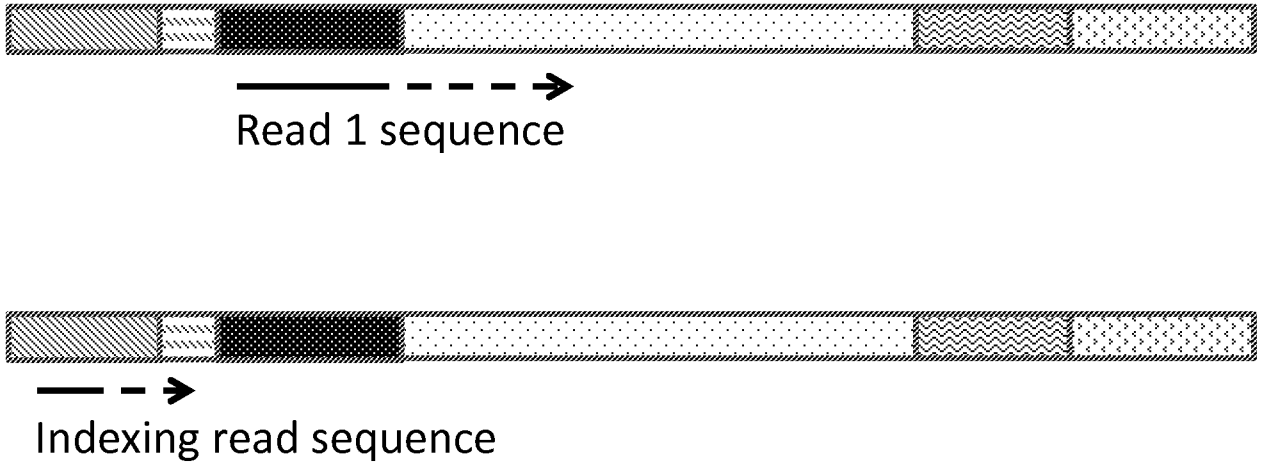
Signature	/Angie Mah/			Date (YYYY-MM-DD)	2014-11-13
First Name	Angie	Last Name	Mah	Registration Number	64902

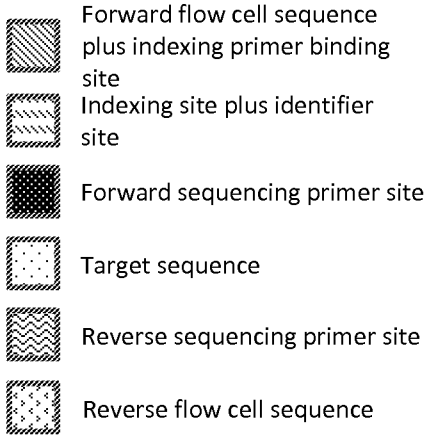
Additional Signature may be generated within this form by selecting the Add button.

This collection of information is required by 37 CFR 1.76. 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 23 minutes to complete, including gathering, preparing, and submitting the completed application data sheet 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.**

Standard DNA Sequencing With Duplication Counter

- Sequence Read 1 (forward read)





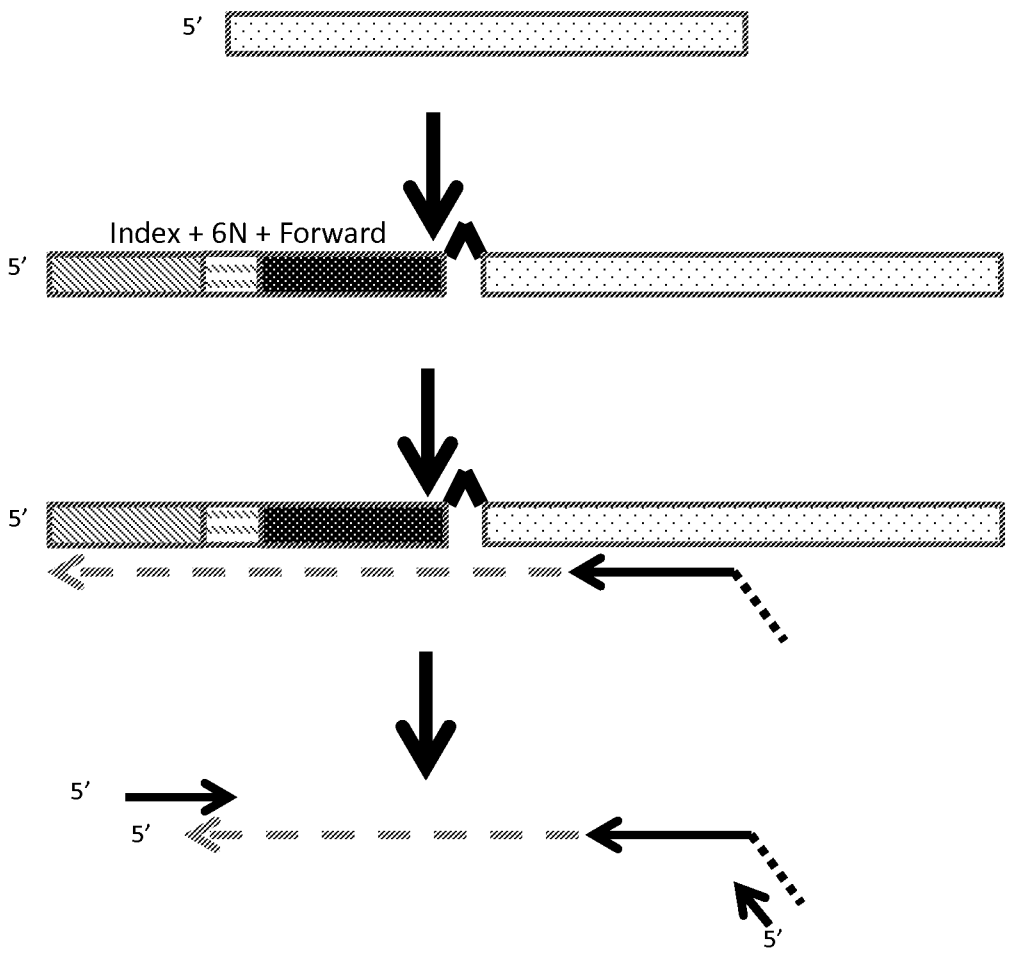
The indexing read results in sequencing of both the index site and the identifier site

- Extension of the indexing read beyond the indexing site and into the nucleotides of the identifier site provides data on which nucleic acid molecule is being sequenced.

Figure 1

IDT Exhibit 1018.0010

Single Primer Extension (SPET) Reaction



gDNA is sheared

Adapters are ligated on to end-repaired DNA. The adapters contain at least a partial indexing sequence priming site, an indexing site, an identifier site, and the forward sequencing priming site. Each gDNA is now labeled with a unique identifier sequence
Targeting oligos are hybridized to gDNA and extended through the adapter containing an identifier site (6N) sequence

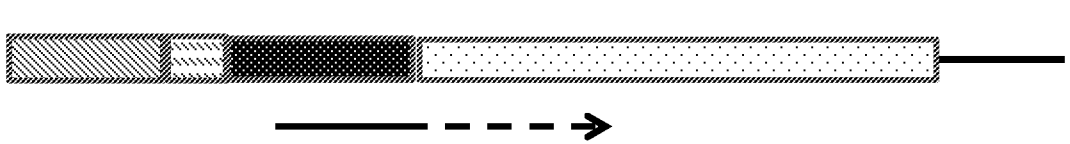
Probes that have extended through an adapter are PCR amplified for sequencing

Figure 2A

IDT Exhibit 1018.0011

SPET Sequencing

- Sequence from random end toward the probe



Read 1 sequence

PCR adds sequence specific primer binding sites and Read 1 of sequencing is performed



Indexing read sequence

Indexing Read performed sequences both the indexing site as well as the identifier site (N6)

- Extending the indexing read past the 6 base indexing site sequence and into the random Ns of the identifier site provides data on which gDNA molecule the extension reaction occurred upon.

Figure 2B

IDT Exhibit 1018.0012

Example Adaptor

An adaptor sequence with necessary priming sites, the index site, and the identifier site

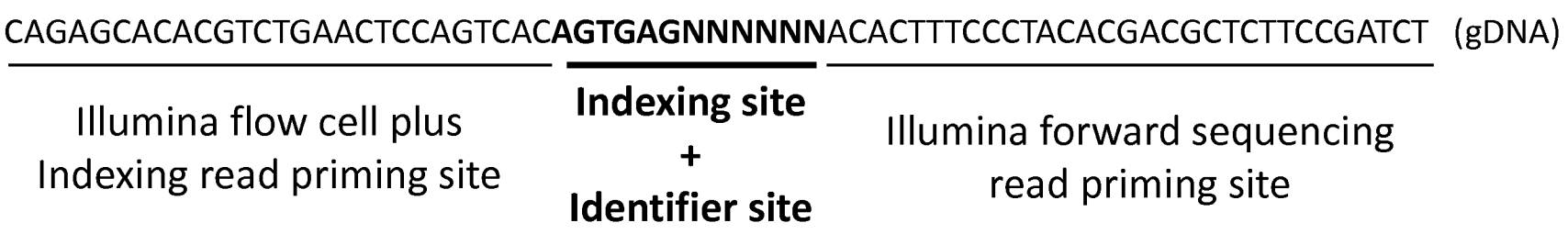
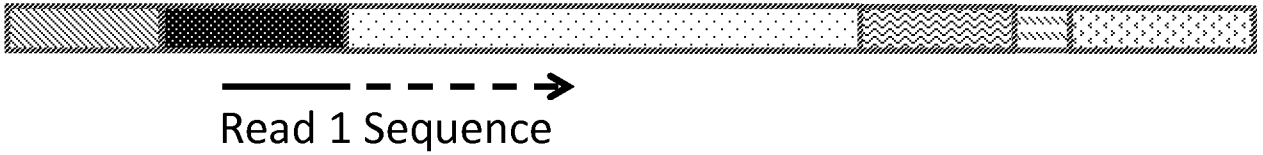


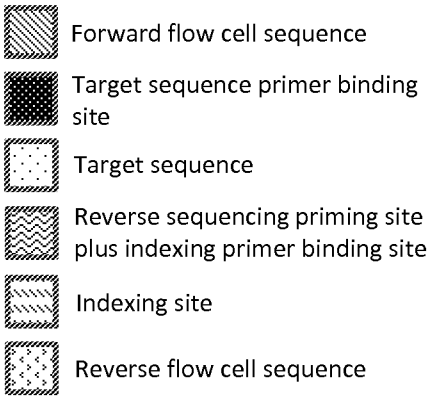
Figure 3

IDT Exhibit 1018.0013

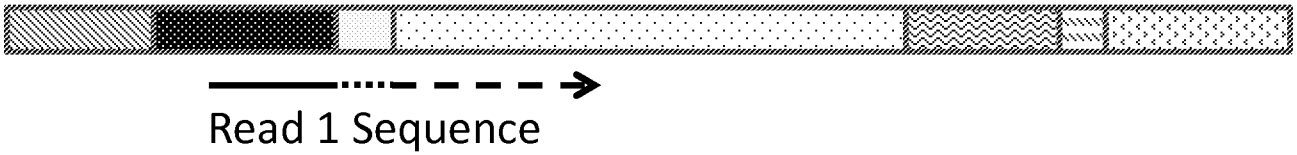
DNA Sequencing With Duplicate Recognition at the insert junction

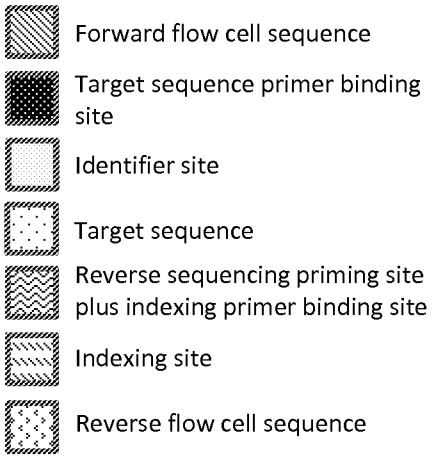
- Traditional library, sequence read 1 (forward read)





- Duplicate recognition library, sequence read 1 (forward read)





Read 1 initially records the approximately 8 random bases of the identifier site appended to the library sequences, followed by the target sequence

Figure 4

IDT Exhibit 1018.0014

Resolution of apparent duplicate reads by identifier sequence ligation

Libraries were generated from 100ng genomic DNA Covaris fragmented to ~500bp in length. These were enriched with the 344 gene cancer panel, ~12k probes covering 969kb of genomic space and sequenced on the ILMN Miseq. Different numbers of reads were randomly selected for this analysis

Total reads randomly selected	Aligned read	Fold Target Coverage	Total duplicates after identifier analysis (true rate)	Percentage duplicates after identifier analysis (true rate)	Total apparent duplicates with no identifier	Percentage apparent duplicates with no identifier
100K	97,919	2X	1854	1.89	3256	3.33
500K	488,423	7X	39731	8.13	68741	14.07
1M	976,662	13X	145589	14.91	237882	24.36
2M	1,954,673	21X	514848	26.34	764825	39.13
10M	9,771,462	43X	6493577	66.45	7515496	76.91

"Apparent" Duplicates – individual targets within libraries consist of one random end (result of fragmentation) and one common end (initiated with the sequence specific probe). Apparent duplicates are derived from sequencing libraries and using conventional measures of duplicate reads ie, reads were mapped using bowtie and all reads with the same start and end genomic coordinates were counted as duplicates.

"True" duplicates - An identifier sequence is introduced through ligation to differentiate between fragments of DNA that randomly have the same start and end mapping coordinates. The analysis method is described in the specification.

Figure 5

IDT Exhibit 1018.0015

COMPOSITIONS AND METHODS FOR IDENTIFICATION OF A DUPLICATE SEQUENCING READ

CROSS-REFERENCE TO RELATED APPLICATIONS

[0001] This application claims the benefit of U.S. Provisional Application Serial No. 61/903,826, filed November 13, 2013, which is herein incorporated by reference in its entirety.

DESCRIPTION OF THE TEXT FILE SUBMITTED ELECTRONICALLY

[0002] The contents of the text file submitted electronically herewith are incorporated herein by reference in their entirety: A computer readable format copy of the Sequence Listing (filename: NUGN_001_01US_SeqList_ST25.txt, date recorded: November 12, 2014, file size 3 kilobytes).

FIELD OF THE PRESENT INVENTION

[0003] The present invention relates generally to the field of high throughput sequencing reactions and the ability to discern artifacts arising through sequence duplications from nucleotide molecules that are unique molecules.

BACKGROUND OF THE PRESENT INVENTION

[0004] In RNA sequencing applications, accurate gene expression measurements may be hampered by PCR duplicate artifacts that occur during library amplification. When analyzing RNA sequencing data, when two or more identical sequences are found, it can be difficult to know if these represent unique cDNA molecules derived independently from different RNA molecules, or if they are PCR duplicates derived from a single RNA molecule. In genotyping by

sequencing, duplicate reads can be considered non-informative and may be collapsed down to a single read, thus reducing the number of sequencing reads used in final analysis. Generally, sequencing reads may be determined to be duplicates if both forward and reverse reads have identical starting positions, even though two independently generated molecules can have identical starting positions by random chance. Single primer extension based targeted re-sequencing suffers from an issue in that only one end of a sequencing read is randomly generated, while the other (reverse read) end is generated by a specific probe. This may make it difficult to determine if two reads are duplicates because they have been duplicated by PCR or because by chance they happened to start at the same position.

[0005] In expression analysis studies there may be limited value in doing paired end sequencing since the goal of the experiment is to determine amounts of transcript present as opposed to studying exon usage. In these studies, paired end sequencing adds costs while the only value is in helping distinguish PCR duplicates. The probability of two reads starting in the same position on only one end is higher than the probability of two reads having the same starting position on two ends (forward and reverse read). There is a need for improved methods that allow for low-cost, high throughput sequencing of regions of interest, genotyping or simple detection of RNA transcripts without inherent instrument inefficiencies that drive up sequencing costs due to the generation of unusable or non-desired data reads. The invention described herein fulfills this need. Here, we describe an adaptor approach that allows for the identification of true PCR duplicates and their removal.

[0006] The methods of the present invention provide novel methods for identifying true duplicate reads during sequencing, such as to improve data analysis of sequencing data, and other related advantages.

SUMMARY OF THE PRESENT INVENTION

[0007] The present invention is based, in part, on compositions and methods for discerning duplicate sequencing reads from a population of sequencing reads. The detection and/or removal of duplicate sequencing reads presented herein is a novel approach to increasing the efficacy of evaluating data generated from high throughput sequence reactions, including complex multiplex sequence reactions.

[0008] Accordingly, the present invention provides a method of detecting a duplicate sequencing read from a population of sample sequencing reads, the method comprising ligating an adaptor to a 5' end of each nucleic acid fragment of a plurality of nucleic acid fragments from one or more samples, wherein the adaptor comprises an indexing primer binding site, an indexing site, an identifier site, and a target sequence primer binding site. The ligated adaptor-nucleic acid fragment products can be amplified, thus generating a population of sequencing reads from the amplified adaptor-nucleic acid ligation products. The sequencing reads with a duplicate identifier site and target sequence can then be detected from the population of sequencing reads. The methods can further include the removal of the sequencing reads with the duplicate identifier site and target sequence from the population of sequence reads.

[0009] In some embodiments, the identifier site is sequenced with the indexing site or the target sequence. In further embodiments, the identifier site is sequenced separately from the indexing site or the target sequence.

[0010] In some embodiments, the adaptor comprises from 5' to 3' the indexing primer binding site; the indexing site; the identifier site; and the target sequence primer binding site. In further embodiments, the adaptor comprises from 5' to 3' the indexing primer binding site; the indexing site; the target sequence primer binding site; and the identifier site.

[0011] In some embodiments, the plurality of nucleic acid fragments is generated from more than one sample. In some embodiments, the nucleic acid fragments from each sample have the same indexing site. In some embodiments, the sequencing reads are separated based on the indexing site. In yet other embodiments, the separation of sequencing reads is performed prior to detecting sequence reads with a duplicate identifier site and target sequence.

[0012] In some embodiments, the nucleic acid fragments are DNA fragments, RNA fragments, or DNA/RNA fragments. In further embodiments, the nucleic acid fragments are genomic DNA fragments or cDNA fragments.

[0013] In some embodiments, the indexing site is between 2 and 8 nucleotides in length. In further embodiments, the indexing site is about 6 nucleotides in length. In some embodiments, the identifier site is between 1 and 8 nucleotides in length. In further embodiments, the identifier site is about 8 nucleotides in length.

[0014] In some embodiments, the indexing primer binding site is a universal indexing primer binding site; and in some embodiments, the target sequence primer binding site is a universal target sequence primer binding site.

[0015] The present invention also encompasses embodiments that include a kit comprising a plurality of adaptors, wherein each adaptor comprise an indexing primer binding site; an indexing site, and identifier site, and a target sequencing primer binding site.

BRIEF DESCRIPTION OF THE DRAWINGS

[0016] A better understanding of the novel features of the invention and advantages of the present invention will be obtained by reference to the following description that sets forth

illustrative embodiments, in which the principles of the invention are utilized, and the accompanying drawings of which:

[0017] Figure 1 depicts a schematic of generating sequencing reads of a library including where an indexing primer and a target sequence primer anneal.

[0018] Figure 2A depicts a mechanism of the Single Primer Enrichment Technology (SPET) and how an identifier site is carried over into the final library and how sequencing through the indexing site into the identifier site provides data on which nucleic acid molecule is being identified.

[0019] Figure 2B is a continuation of Figure 2A.

[0020] Figure 3 offers a detailed view of the sequence of an example adaptor, among many envisioned embodiments, and the position of the indexing and identifier sites (SEQ ID NO: 1). "N" refers to any nucleic acid.

[0021] Figure 4 depicts a schematic of two separate sequence libraries, indicating where the indexing primers and the target primers anneal in a traditional library as compared to a library using an identifier site.

[0022] Figure 5 depicts a data table demonstrating the accuracy of resolving true duplicates versus apparent or perceived duplicates using an identifier site in the adaptors.

DETAILED DESCRIPTION OF THE PRESENT INVENTION

[0023] The present invention is based, in part, on compositions and methods for discerning duplicate sequencing reads from a population of sequencing reads. The present invention encompasses methods of detecting sequences duplicated in sequencing applications, and further removal of the duplicated sequence reads. The present invention further encompasses kits

comprising components that would allow for customized applications of the method of detecting and removing duplicated sequence reads in high throughput sequencing reactions. The compositions and methods can be used with various applications for genetic sample analysis, such as RNA sequence analysis, copy number variation analysis, methylation sequencing analysis, genotyping and whole genome amplification.

[0024] Reference will now be made in detail to exemplary embodiments of the invention. While the disclosed methods and compositions will be described in conjunction with the exemplary embodiments, it will be understood that these exemplary embodiments are not intended to limit the present invention. On the contrary, the present disclosure is intended to encompass alternatives, modifications and equivalents, which may be included in the spirit and scope of the present invention.

[0025] Unless otherwise specified, terms and symbols of genetics, molecular biology, biochemistry and nucleic acid used herein follow those of standard treatises and texts in the field, e.g. Kornberg and Baker, *DNA Replication,* Second Edition (W.H. Freeman, New York, 1992); Lehninger, *Biochemistry,* Second Edition (Worth Publishers, New York, 1975); Strachan and Read, Human *Molecular Genetics,* Second Edition (Wiley-Liss, New York, 1999); Eckstein, editor, *Oligonucleotides and Analogs: A Practical Approach* (Oxford University Press, New York, 1991); Gait, editor, *Oligonucleotide Synthesis: A Practical Approach* (IRL Press, Oxford,1984); and the like.

[0026] In some embodiments, the methods disclosed herein is for detecting from a population of sequencing reads a sequencing read, such as a duplicate sequencing read with a duplicate identifier site and target sequence. A duplicate sequencing read can be a sequencing read with

the same identifier site and target sequence as another sequencing read in the population of sequencing reads.

Adaptor

[0027] The present invention provides compositions of adaptors and methods comprising use of an adaptor. An adapter refers to an oligonucleotide sequence, the ligation of which to a target polynucleotide or a target polynucleotide strand of interest enables the generation of amplification-ready products of the target polynucleotide or the target polynucleotide strand of interest. The target polynucleotide molecules may be fragmented or not prior to the addition of adaptors. In some embodiments, a method disclosed herein comprises ligating an adaptor to a 5' end of each nucleic acid fragment of a plurality of nucleic acid fragments from one or more samples.

[0028] Various adaptor designs are envisioned which are suitable for generation of amplification-ready products of target sequence regions/strands of interest. For example, the two strands of the adaptor may be self-complementary, non-complementary or partially complementary. In some embodiments, the adaptor can comprise an indexing primer binding site, an indexing site, an identifier site, and a target sequence primer binding site.

[0029] An indexing primer binding site is a nucleotide sequence for binding a primer for an indexing site. An indexing site is a nucleic acid sequence that acts as an index for multiple polynucleotide samples, thus allowing for the samples to be pooled together into a single sequencing run, which is known as multiplexing. In some embodiments, the indexing site is at least 2, 3, 4, 5, 6, 7, 8, 9, or 10 nucleotides in length. In some embodiments, the indexing site is between 2 and 8 nucleotides in length. In some embodiments, the indexing site is about 6 nucleotides in length.

[0030] An identifier site is a nucleic acid sequence that comprises random bases and is used to identify duplicate sequencing reads. In some embodiments, the identifier site is at least 2, 3, 4, 5, 6, 7, 8, 9, or 10 nucleotides in length. In some embodiments, the identifier site is between 1 and 8 nucleotides in length. In some embodiments, the identifier site is about 8 nucleotides in length. This identifier site can be designed as a set of sequences, or it can be semi-random, or it can be completely random. In addition, this identifier site can be a fixed length, or it can be a variable length. In some embodiments, the identifier sites in a plurality of adaptors are of a fixed length. For example, the identifier sites can all be eight random bases. In another embodiment, the identifier sites in a plurality of adaptors are of a variable length. For example, the identifier sites can range in size from 1 to 8 bases. In yet another embodiment, the identifier sites can be of a defined set of defined sequence. For example, the identifier site of a plurality of adaptors can be one of 96 defined six-base nucleotide sequences.

[0031] A target sequence primer binding site nucleotide sequence for binding a primer for a target sequence. The primer can be used to amplify the target sequence (e.g., a nucleic acid fragment from a sample). Accordingly, in some embodiments, the adaptor comprises an indexing primer binding site and a target sequence primer binding site.

[0032] A primer is a polynucleotide chain, typically less than 200 residues long, most typically between 15 and 100 nucleotides long, but can encompass longer polynucleotide chains. A primer targeting the primer binding sites are typically designed to hybridize to single-stranded nucleic acid strands. In some embodiments, the primers targeting the primer binding sites are designed to hybridize to single-stranded DNA targets. In the case where the sample comprises genomic DNA or other double-stranded DNA, the sample can be first denatured to render the target single stranded and enable hybridization of the primers to the desired sequence regions of

interest. In these embodiments, the methods and compositions described herein can allow for region-specific enrichment and amplification of sequence regions of interest. In some embodiments, the other double-stranded DNA can be double-stranded cDNA generated by first and second strand synthesis of one or more target RNAs.

[0033] In other embodiments, the primers targeting the primer binding sites are designed to hybridize to double-stranded nucleic acid targets, without denaturation of the double stranded nucleic acids. In other embodiments, the primers targeting the primer binding sites are designed to hybridize to a double-stranded DNA target, without denaturation of the dsDNA. In these embodiments, the primers targeting the selected sequence regions of interest are designed to form a triple helix (triplex) at the selected sequence regions of interest. The hybridization of the primers to the double-stranded DNA sequence regions of interest can be carried out without prior denaturation of the double stranded nucleic acid sample. In such embodiments, the methods and compositions described herein can allow for region-specific enrichment as well as strand-specific enrichment and amplification of sequence regions of interest. This method can be useful for generation of copies of strand specific sequence regions of interest from complex nucleic acid without the need to denature the dsDNA input DNA, thus enabling enrichment and analysis of multiplicity of sequence regions of interest in the native complex nucleic acid sample. The method can find use for studies and analyses carried out in situ, enable studies and analysis of complex genomic DNA in single cells or collection of very small well defined cell population, as well as permit the analysis of complex genomic DNA without disruption of chromatin structures.

[0034] The primers of the invention are generally oligonucleotides that are employed in an extension reaction by a polymerase along a polynucleotide template, such as for amplification of a target sequence (e.g., in PCR). The oligonucleotide primer can be a synthetic polynucleotide

that is single stranded, containing a sequence at its 3'-end that is capable of hybridizing with a sequence of the target polynucleotide. In some embodiments, the 3' region of the primer that hybridizes with the target nucleic acid has at least 80%, preferably 90%, more preferably 95%, most preferably 100%, complementarity to a primer binding site.

[0035] In some embodiments, the primer binding site is a binding site for a universal primer. A universal primer is a primer that can be used for amplifying a number of different sequences. In some embodiments, a universal primer is used to amplify different libraries. In some embodiments, an indexing primer binding site is a binding site for a universal indexing primer (i.e., the indexing primer binding site is a universal indexing primer binding site). In some embodiments, the adaptors used for ligating a plurality of nucleic acid fragments have a universal indexing primer binding site. In some embodiments, the universal indexing primer can be used to amplify and/or sequence a number of different indexing sites.

[0036] In some embodiments, a target sequence primer binding site is a binding site for a universal target sequence primer (i.e., the target sequence primer binding site is a universal target sequence primer binding site). In some embodiments, the adaptors used for ligating a plurality of nucleic acid fragments have a universal target sequence primer binding site. In some embodiments, the universal target sequence primer can be used to amplify and/or sequence a number of different target sequences.

[0037] In some embodiments, the adaptor comprises an identifier site 3' to an indexing site. In some embodiments, the adaptor comprises an identifier site 5' to an indexing site. In some embodiments, the adaptor comprises from 5' to 3' an indexing primer binding site, indexing site, identifier site, and target sequence primer binding site. In other embodiments, the adaptor comprises from 5' to 3' an indexing primer binding site, indexing site, identifier site, and target

sequence primer binding site. In yet other embodiments, the adaptor comprises from 5' to 3' an indexing primer binding site, identifier site, indexing site, and target sequence primer binding site.

Samples

[0038] In some embodiments, an adaptor is ligated to a nucleic acid fragment (e.g., the 5' end of the nucleic acid fragment). The nucleic acid fragment can be from a plurality of nucleic acid fragments from one or more samples. The nucleic acid fragment can be RNA, DNA, or complex DNA, for example genomic DNA and PNA, in which case one might use a modified nucleic acid. The nucleic acid fragment may also be cDNA. The cDNA can be generated from RNA, e.g., mRNA.

[0039] The sample can be a biological sample. For example, the sample can be an animal, plant, bacterial, algal, or viral sample. In some embodiments, the sample is a human, rat, or mouse sample. The sample can be from a mixture of genomes of different species such as host-pathogen, bacterial populations and the like. The sample can be cDNA made from a mixture of genomes of different species. In some embodiments, the sample can be from a synthetic source. The sample can be mitochondrial DNA. The sample can be cell-free DNA. The cell-free DNA can be obtained from sources such as a serum or a plasma sample. The sample can comprise one or more chromosomes. For example, if the sample is from a human, the sample can comprise one or more of chromosome 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, X, or Y. In some embodiments, the sample comprises a linear or circular genome. The sample can be plasmid DNA, cosmid DNA, bacterial artificial chromosome (BAC), or yeast artificial chromosome (YAC). The sample can be from more than one individual or organism. The sample

can be double-stranded or single-stranded. The sample can be part of chromatin. The sample can be associated with histones.

[0040] In some embodiments, adaptors are ligated to a plurality of nucleic acid fragments from more than one sample, such as 2, 3, 4, 5, or more samples. In some embodiments, the nucleic acid fragments from each sample have the same indexing site. In some embodiments, a plurality of nucleic acid fragments is generated from a first sample and a second sample and adaptors are ligated to each nucleic acid fragment, in which adaptors ligated to each nucleic acid fragment from the first sample have the same first indexing site and adaptors ligated to nucleic acid fragments from the second sample has the same second indexing site. In some embodiments, the nucleic acid fragments or data associated with the nucleic acid fragments (e.g., sequencing reads) are separated based on the indexing site.

[0041] In some embodiments, a population of nucleic acid fragments generated from a sample is of one or more specific size range(s). In some embodiments, the fragments have an average length from about 10 to about 10,000 nucleotides. In some embodiments, the fragments have an average length from about 50 to about 2,000 nucleotides. In some embodiments, the fragments have an average length from about 100-2,500, 10-1,000, 10-800, 10-500, 50-500, 50-250, or 50-150 nucleotides. In some embodiments, the fragments have an average length less than 10,000 nucleotide, such as less than 5,000 nucleotides, less than 2,500 nucleotides, less than 2,500 nucleotides, less than 1,000 nucleotides, less than 500 nucleotides, such as less than 400 nucleotides, less than 300 nucleotides, less than 200 nucleotides, or less than 150 nucleotides.

[0042] In some embodiments, fragmentation of the nucleic acids can be achieved through methods known in the art. Fragmentation can be achieved through physical fragmentation methods and/or enzymatic fragmentation methods. Physical fragmentation methods can include

nebulization, sonication, and/or hydrodynamic shearing. In some embodiments, the fragmentation can be accomplished mechanically comprising subjecting the nucleic acids in the input sample to acoustic sonication. In some embodiments, the fragmentation comprises treating the nucleic acids in the input sample with one or more enzymes under conditions suitable for the one or more enzymes to generate double-stranded nucleic acid breaks. Examples of enzymes useful in the generation of nucleic acid or polynucleotide fragments include sequence specific and non-sequence specific nucleases. Non-limiting examples of nucleases include DNase I, Fragmentase, restriction endonucleases, variants thereof, and combinations thereof. Reagents for carrying out enzymatic fragmentation reactions are commercially available (e.g., from New England Biolabs). For example, digestion with DNase I can induce random double-stranded breaks in DNA in the absence of Mg^{++} and in the presence of Mn^{++}. In some embodiments, fragmentation comprises treating the nucleic acids in the input sample with one or more restriction endonucleases. Fragmentation can produce fragments having 5′ overhangs, 3′ overhangs, blunt ends, or a combination thereof. In some embodiments, such as when fragmentation comprises the use of one or more restriction endonucleases, cleavage of sample polynucleotides leaves overhangs having a predictable sequence. In some embodiments, the method includes the step of size selecting the fragments via standard methods known in the art such as column purification or isolation from an agarose gel.

[0043] In some embodiments, fragmentation of the nucleic acids is followed by end repair of the nucleic acid fragments. End repair can include the generation of blunt ends, non-blunt ends (i.e sticky or cohesive ends), or single base overhangs such as the addition of a single dA nucleotide to the 3′-end of the nucleic acid fragments, by a polymerase lacking 3′-exonuclease activity. End repair can be performed using any number of enzymes and/or methods known in the art

including, but not limited to, commercially available kits such as the Encore™ Ultra Low Input NGS Library System I. In some embodiments, end repair can be performed on double stranded DNA fragments to produce blunt ends wherein the double stranded DNA fragments contain 5′ phosphates and 3′ hydroxyls. In some embodiments, the double-stranded DNA fragments can be blunt-end polished (or "end-repaired") to produce DNA fragments having blunt ends, prior to being joined to adapters. Generation of the blunt ends on the double stranded fragments can be generated by the use of a single strand specific DNA exonuclease such as for example exonuclease 1, exonuclease 7 or a combination thereof to degrade overhanging single stranded ends of the double stranded products. Alternatively, the double stranded DNA fragments can be blunt ended by the use of a single stranded specific DNA endonuclease, for example, but not limited to, mung bean endonuclease or S1 endonuclease. Alternatively, the double stranded products can be blunt ended by the use of a polymerase that comprises single stranded exonuclease activity such as for example T4 DNA polymerase, or any other polymerase comprising single stranded exonuclease activity or a combination thereof to degrade the overhanging single stranded ends of the double stranded products. In some cases, the polymerase comprising single stranded exonuclease activity can be incubated in a reaction mixture that does or does not comprise one or more dNTPs. In other cases, a combination of single stranded nucleic acid specific exonucleases and one or more polymerases can be used to blunt end the double stranded fragments generated by fragmenting the sample comprising nucleic acids. In still other cases, the nucleic acid fragments can be made blunt ended by filling in the overhanging single stranded ends of the double stranded fragments. For example, the fragments may be incubated with a polymerase such as T4 DNA polymerase or Klenow polymerase or a combination thereof in the presence of one or more dNTPs to fill in the single stranded portions

of the double stranded fragments. Alternatively, the double stranded DNA fragments can be made blunt by a combination of a single stranded overhang degradation reaction using exonucleases and/or polymerases, and a fill-in reaction using one or more polymerases in the presence of one or more dNTPs.

[0044] U.S. Patent Publication Nos. 2013-0231253 A1 and 2014-0274729 A1 further describe methods of generating nucleic acid fragments, methods of modifying the fragments and analysis of the fragments, and are incorporated by reference in their entirety herein.

Ligation of Adaptors

[0045] Ligation of adaptors at the desired end of the sequence regions of interest (e.g., at the 5' or 3' end of a nucleic acid fragment generated form a sample) is suitable for carrying out the methods of the invention. Various ligation modalities are envisioned, dependent on the choice of nucleic acid, nucleic acid modifying enzymes and the resulting ligatable end of the nucleic acid. For example, when a blunt end product comprising the target region/sequence of interest is generated, blunt end ligation can be suitable. Alternatively, where the cleavage is carried out using a restriction enzyme of known sequence specificity, leading to the generation of cleavage sites with known sequence overhangs, suitable ends of the adaptors can be designed to enable hybridization of the adaptor to the cleavage site of the sequence region of interest and subsequent ligation. Ligation also refers to any joining of two nucleic acid molecules that results in a single nucleic acid sequence that can be further modified to obtain the sequence of the nucleic acids in question. Reagents and methods for efficient and rapid ligation of adaptors are commercially available, and are known in the art.

[0046] In some embodiments, the 5′ and/or 3′ end nucleotide sequences of fragmented nucleic acids are not modified or end-repaired prior to ligation with the adapter oligonucleotides of the present invention. For example, fragmentation by a restriction endonuclease can be used to leave

a predictable overhang, followed by ligation with one or more adapter oligonucleotides comprising an overhang complementary to the predictable overhang on a nucleic acid fragment. In another example, cleavage by an enzyme that leaves a predictable blunt end can be followed by ligation of blunt-ended nucleic acid fragments to adapter oligonucleotides comprising a blunt end. In some embodiments, end repair can be followed by an addition of 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20 or more nucleotides, such as one or more adenine, one or more thymine, one or more guanine, or one or more cytosine, to produce an overhang. Nucleic acid fragments having an overhang can be joined to one or more adapter oligonucleotides having a complementary overhang, such as in a ligation reaction. For example, a single adenine can be added to the 3′ ends of end-repaired DNA fragments using a template independent polymerase, followed by ligation to one or more adapters each having a thymine at a 3′ end. In some embodiments, adapter oligonucleotides can be joined to blunt end double-stranded nucleic acid fragments which have been modified by extension of the 3′ end with one or more nucleotides followed by 5′ phosphorylation. In some cases, extension of the 3′ end can be performed with a polymerase such as for example Klenow polymerase or any of the suitable polymerases provided herein, or by use of a terminal deoxynucleotide transferase, in the presence of one or more dNTPs in a suitable buffer containing magnesium. In some embodiments, nucleic acid fragments having blunt ends can be joined to one or more adapters comprising a blunt end. Phosphorylation of 5′ ends of nucleic acid fragments can be performed for example with T4 polynucleotide kinase in a suitable buffer containing ATP and magnesium. The fragmented nucleic acid molecules may optionally be treated to dephosphorylate 5′ ends or 3′ ends, for example, by using enzymes known in the art, such as phosphatases.

[0047] In some embodiments, appending the adaptor to the nucleic acid fragments generated by methods described herein can be achieved using a ligation reaction or a priming reaction. In some embodiments, appendage of an adaptor to the nucleic acid fragments comprises ligation. In some embodiments, ligation of the adaptor to the nucleic acid fragments can be following end repair of the nucleic acid fragments. In another embodiment, the ligation of the adaptor to the nucleic acid fragments can be following generation of the nucleic acid fragments without end repair of the nucleic acid fragments. The adaptor can be any type of adaptor known in the art including, but not limited to, conventional duplex or double stranded adaptors in which the adaptor comprises two complementary strands. In some embodiments, the adaptor can be a double stranded DNA adaptor. In some embodiments, the adaptor can be an oligonucleotide of known sequence and, thus, allow generation and/or use of sequence specific primers for amplification and/or sequencing of any polynucleotides to which the adaptor is appended or attached. In some embodiments, the adaptor can be a conventional duplex adaptor, wherein the adaptor comprises sequence well known in the art. In a some embodiments, the adaptor can be appended to the nucleic acid fragments generated by the methods described herein in multiple orientations. In some embodiment, the methods described herein can involve the use of a duplex adaptor comprising double stranded DNA of known sequence that is blunt ended and can bind to the double stranded nucleic acid fragments generated by the methods described herein in one of two orientations. In some embodiments, the adaptor can be ligated to each of the nucleic acid fragments such that each of the nucleic acid fragments comprises the same adaptor. In other words, each of the nucleic acid fragments comprises a common adaptor. In another embodiment, an adaptor can be appended or ligated to a library of nucleic acid fragments generated by the methods described herein such that each nucleic acid fragment in the library of nucleic acid

fragments comprises the adaptor ligated to one or both ends. In another embodiment, more than one adaptor can be appended or ligated to a library of nucleic acid fragments generated by the methods described herein. The multiple adaptors may occur adjacent to one another, spaced intermittently, or at opposite ends of the nucleic acid fragments. In some embodiments, the adaptor can be ligated or appended to the 5′ and/or 3′ ends of the nucleic acid fragments generated by the methods described herein. The adaptor can comprise two strands wherein each strand comprises a free 3′ hydroxyl group but neither strand comprises a free 5′ phosphate. In some embodiments, the free 3′ hydroxyl group on each strand of the adaptor can be ligated to a free 5′ phosphate present on either end of the nucleic acid fragments of the present invention. In this embodiment, the adaptor comprises a ligation strand and a non-ligation strand whereby the ligation strand can be ligated to the 5′ phosphate on either end of the nucleic acid fragment while a nick or gap can be present between the non-ligation strand of the adaptor and the 3′ hydroxyl on either end of the nucleic acid fragment. In some embodiments, the nick or gap can be filled in by performing a gap repair reaction. In some embodiments, the gap repair can be performed with a DNA dependent DNA polymerase with strand displacement activity. In some embodiments, the gap repair can be performed using a DNA-dependent DNA polymerase with weak or no strand displacement activity. In some embodiments, the ligation strand of the adaptor can serve as the template for the gap repair or fill-in reaction. The gap repair or fill-in reaction may comprise an extension reaction wherein the ligation strand of the adaptor serves as a template and leads to the generation of nucleic acid fragments with complementary termini or ends. In some embodiments, the gap repair can be performed using Taq DNA polymerase. In some embodiments, the ligation of the first adaptor to the nucleic acid fragments generated by the

methods described herein may not be followed by gap repair. The nucleic acid fragments may comprise the adaptor sequence ligated only at the 5′ end of each strand.

[0048] Ligation and, optionally gap repair, of the adaptor to the nucleic acid fragments generates an adaptor-nucleic acid fragment complex. In some embodiments, the adaptor-nucleic acid fragment complex can be denatured. Denaturation can be achieved using any of the methods known in the art including, but not limited to, physical, thermal, and/or chemical denaturation. In some embodiments, denaturation can be achieved using thermal or heat denaturation. In some embodiments, denaturation of the adaptor-nucleic acid fragment complex generates single stranded nucleic acid fragments comprising the adaptor sequence at only the 5′end of the nucleic acid fragments. In another embodiment, denaturation of the first adaptor-nucleic acid fragment complex generates single stranded nucleic acid fragments comprising adaptor sequence at both the 5′end and 3′end of the nucleic acid fragments.

Methods of Amplification

[0049] The methods, compositions and kits described herein can be useful to generate amplification-ready products directly from a nucleic acid source for downstream applications such as next generation sequencing, as well as generation of libraries with enriched population of sequence regions of interest. In some embodiments, the adapter-nucleic fragment ligated products, e.g., from the ligation of an adaptor to a 5’ end of each nucleic acid fragment of a plurality of nucleic acid fragments from one or more samples, is amplified.

[0050] Methods of amplification are well known in the art. In some embodiments, the amplification is exponential, e.g. in the enzymatic amplification of specific double stranded sequences of DNA by a polymerase chain reaction (PCR). In other embodiments the amplification method is linear. In other embodiments the amplification method is isothermal. In

some embodiments, the amplification is exponential, e.g. in the enzymatic amplification of specific double stranded sequences of DNA by a polymerase chain reaction (PCR).

[0051] Suitable amplification reactions can be exponential or isothermal and can include any DNA amplification reaction, including but not limited to polymerase chain reaction (PCR), strand displacement amplification (SDA), linear amplification, multiple displacement amplification (MDA), rolling circle amplification (RCA), single primer isothermal amplification (SPIA, see e.g. U.S. Pat. No. 6,251,639), Ribo-SPIA, or a combination thereof. In some cases, the amplification methods for providing the template nucleic acid may be performed under limiting conditions such that only a few rounds of amplification (e.g., 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30 etc.), such as for example as is commonly done for cDNA generation. The number of rounds of amplification can be about 1-30, 1-20, 1-15, 1-10, 5-30, 10-30, 15-30, 20-30, 10-30, 15-30, 20-30, or 25-30.

[0052] PCR is an in vitro amplification procedure based on repeated cycles of denaturation, oligonucleotide primer annealing, and primer extension by thermophilic template dependent polynucleotide polymerase, resulting in the exponential increase in copies of the desired sequence of the polynucleotide analyte flanked by the primers. The two different PCR primers, which anneal to opposite strands of the DNA, are positioned so that the polymerase catalyzed extension product of one primer can serve as a template strand for the other, leading to the accumulation of a discrete double stranded fragment whose length is defined by the distance between the 5′ ends of the oligonucleotide primers.

[0053] LCR uses a ligase enzyme to join pairs of preformed nucleic acid probes. The probes hybridize with each complementary strand of the nucleic acid analyte, if present, and ligase is

employed to bind each pair of probes together resulting in two templates that can serve in the next cycle to reiterate the particular nucleic acid sequence.

[0054] SDA (Westin et a 2000, Nature Biotechnology, 18, 199-202; Walker et al 1992, Nucleic Acids Research, 20, 7, 1691-1696), is an isothermal amplification technique based upon the ability of a restriction endonuclease such as HincII or BsoBI to nick the unmodified strand of a hemiphosphorothioate form of its recognition site, and the ability of an exonuclease deficient DNA polymerase such as Klenow exo minus polymerase, or Bst polymerase, to extend the 3′-end at the nick and displace the downstream DNA strand. Exponential amplification results from coupling sense and antisense reactions in which strands displaced from a sense reaction serve as targets for an antisense reaction and vice versa.

[0055] Some aspects of the invention utilize linear amplification of nucleic acids or polynucleotides. Linear amplification generally refers to a method that involves the formation of one or more copies of the complement of only one strand of a nucleic acid or polynucleotide molecule, usually a nucleic acid or polynucleotide analyte. Thus, the primary difference between linear amplification and exponential amplification is that in the latter process, the product serves as substrate for the formation of more product, whereas in the former process the starting sequence is the substrate for the formation of product but the product of the reaction, i.e. the replication of the starting template, is not a substrate for generation of products. In linear amplification the amount of product formed increases as a linear function of time as opposed to exponential amplification where the amount of product formed is an exponential function of time.

Downstream Applications

[0056] One aspect of the present invention is that the methods and compositions disclosed herein can be efficiently and cost-effectively utilized for downstream analyses, such as in next generation sequencing or hybridization platforms, with minimal loss of biological material of interest. The methods of the present invention can also be used in the analysis of genetic information of selective genomic regions of interest (e.g., analysis of SNPs or other disease markers) as well as genomic regions which may interact with the selective region of interest. The methods of the present invention may further be used in the analysis of copy number variation as well as differential expression.

Sequencing

[0057] In some embodiments, a population of sequencing reads is generated from the amplified adapter-nucleic fragment ligated products. In some embodiments, a sequencing read comprises an index read, which comprises the sequence of an indexing site. In some embodiments, an index read comprises the sequence of an indexing site and the sequence of an identifier site. For example, the indexing site is sequenced with the identifier site. In some embodiments, the index read does not include the sequence of an identifier site. For example, the indexing site is not sequenced with the identifier site. In some embodiments, a sequencing read comprises the target sequence. In some embodiments, a sequencing read comprises a target sequence and an identifier sequence. For example, the target sequence is sequenced with the identifier site. In some embodiments, the target sequence is not sequenced with the identifier site.

[0058] The methods of the present invention are useful for sequencing by the method commercialized by Illumina, as described U.S. Pat Nos. 5,750,341; 6,306,597; and 5,969,119.

[0059] In general, double stranded fragment polynucleotides can be prepared by the methods of the present invention to produce amplified nucleic acid sequences tagged at one (e.g., (A)/(A′) or both ends (e.g., (A)/(A′) and (C)/(C′)). In some cases, single stranded nucleic acid tagged at one or both ends is amplified by the methods of the present invention (e.g., by SPIA or linear PCR).The resulting nucleic acid is then denatured and the single-stranded amplified polynucleotides are randomly attached to the inside surface of flow-cell channels. Unlabeled nucleotides are added to initiate solid-phase bridge amplification to produce dense clusters of double-stranded DNA. To initiate the first base sequencing cycle, four labeled reversible terminators, primers, and DNA polymerase are added. After laser excitation, fluorescence from each cluster on the flow cell is imaged. The identity of the first base for each cluster is then recorded. Cycles of sequencing are performed to determine the fragment sequence one base at a time.

[0060] In some embodiments, the methods of the invention are useful for preparing target polynucleotides for sequencing by the sequencing by ligation methods commercialized by Applied Biosystems (e.g., SOLiD sequencing). In other embodiments, the methods are useful for preparing target polynucleotides for sequencing by synthesis using the methods commercialized by 454/Roche Life Sciences, including but not limited to the methods and apparatus described in Margulies et al., *Nature (*2005) 437:376-380 (2005); and U.S. Pat. Nos. 7,244,559; 7,335,762; 7,211,390; 7,244,567; 7,264,929; and 7,323,305. In other embodiments, the methods are useful for preparing target polynucleotide(s) for sequencing by the methods commercialized by Helicos BioSciences Corporation (Cambridge, Mass.) as described in U.S. application Ser. No. 11/167,046, and U.S. Pat. Nos. 7,501,245; 7,491,498; 7,276,720; and in U.S. Patent Application Publication Nos. US20090061439; US20080087826; US20060286566; US20060024711;

US20060024678; US20080213770; and US20080103058. In other embodiments, the methods are useful for preparing target polynucleotide(s) for sequencing by the methods commercialized by Pacific Biosciences as described in U.S. Pat. Nos. 7,462,452; 7,476,504; 7,405,281; 7,170,050; 7,462,468; 7,476,503; 7,315,019; 7,302,146; 7,313,308; and US Application Publication Nos. US20090029385; US20090068655; US20090024331; and US20080206764.

[0061] An example of a sequencing technique that can be used in the methods of the provided invention is semiconductor sequencing provided by Ion Torrent (e.g., using the Ion Personal Genome Machine (PGM)). Ion Torrent technology can use a semiconductor chip with multiple layers, e.g., a layer with micro-machined wells, an ion-sensitive layer, and an ion sensor layer. Nucleic acids can be introduced into the wells, e.g., a clonal population of single nucleic can be attached to a single head, and the bead can be introduced into a well. To initiate sequencing of the nucleic acids on the beads, one type of deoxyribonucleotide (e.g., dATP, dCTP, dGTP, or dTTP) can be introduced into the wells. When one or more nucleotides are incorporated by DNA polymerase, protons (hydrogen ions) are released in the well, which can be detected by the ion sensor. The semiconductor chip can then be washed and the process can be repeated with a different deoxyribonucleotide. A plurality of nucleic acids can be sequenced in the wells of a semiconductor chip. The semiconductor chip can comprise chemical-sensitive field effect transistor (chemFET) arrays to sequence DNA (for example, as described in U.S. Patent Application Publication No. 20090026082). Incorporation of one or more triphosphates into a new nucleic acid strand at the 3' end of the sequencing primer can be detected by a change in current by a chemFET. An array can have multiple chemFET sensors.

[0062] Another example of a sequencing technique that can be used in the methods of the provided invention is nanopore sequencing (see e.g., Soni G V and Meller A. (2007) *Clin Chem*

53: 1996-2001). A nanopore can be a small hole of the order of 1 nanometer in diameter. Immersion of a nanopore in a conducting fluid and application of a potential across it can result in a slight electrical current due to conduction of ions through the nanopore. The amount of current that flows is sensitive to the size of the nanopore. As a DNA molecule passes through a nanopore, each nucleotide on the DNA molecule obstructs the nanopore to a different degree. Thus, the change in the current passing through the nanopore as the DNA molecule passes through the nanopore can represent a reading of the DNA sequence.

Data Analysis

[0063] In some embodiments, the sequence reads are used in the analysis of genetic information of selective genomic regions of interest as well as genomic regions which may interact with the selective region of interest. Amplification methods as disclosed herein can be used in the devices, kits, and methods known to the art for genetic analysis, such as, but not limited to those found in U.S. Pat. Nos. 6,449,562, 6,287,766, 7,361,468, 7,414,117, 6,225,109, and 6,110,709.

[0064] In some embodiments, the sequencing reads are used to detect duplicate sequencing reads. In some embodiments, a sequencing read is identified as a duplicate sequencing read when it contains an identifier site and target sequence that is the same as another sequencing read from the same population of sequencing reads.

[0065] In some embodiments, duplicate sequencing reads are differentiated from one another as being true duplicates versus apparent or perceived duplicates. Apparent or perceived duplicates may be identified from sequencing libraries and using conventional measures of duplicate reads (i.e., reads were mapped using bowtie), where all reads with the same start and end nucleic acid coordinates were counted as duplicates. True duplicates may be identified from sequencing

libraries having had an identifier site introduced through ligation to differentiate between fragments of DNA that randomly have the same start and end mapping coordinates.

[0066] In some embodiments, the sequence of the identifier sites from the index read of two nucleic acids generated from any dsDNA may have the same start site, as determined from the sequencing reads of the target sequence of the generated nucleic acid fragments. If the identifier site from the index read of the two nucleic acid fragments are not identical, then the target sequence reads were not generated from the same original dsDNA molecule and, therefore, are no true duplicate reads. The ligation of random sequences onto dsDNA molecules, accompanied by the methods of the invention, allow for the identification of true duplicate reads versus apparent or perceived duplicate reads.

[0067] In some embodiments, the sequence of the identifier sites from the index read of two nucleic acid fragments generated from a genomic DNA (gDNA) molecule that have the same start site, which can be determined from the sequencing reads of the target sequence of the nucleic acid fragments is determined. If the identifier site from the index read of the two nucleic acid fragments are not identical, then the target sequence reads were not generated from the same original gDNA molecule and, therefore, are not true duplicate reads. In another embodiment, an identifier site is inserted at the adapter insert junction. The sequence of the identifier site is carried through the library amplification step. The identifier site is the first sequence read during the forward read. As the identifier sequence is not logically present adjacent to naturally occurring sequence, this uniquely identifies the DNA fragment. Therefore, by ligating random sequences onto the original gDNA, the methods of the invention identify true duplicate reads.

[0068] In some embodiments, a duplicate sequencing read is detected and analyzed. Duplicate reads can be filtered using 'samtools rmdup', wherein reads with identical external coordinates

are removed, only one read with highest mapping quality is retained. After filtering, a filtered set of deduplicated reads can be used in any downstream analysis. Conversely, this filtering step can be skipped, and downstream analysis can be done using the unfiltered reads, including duplicates.

[0069] In some embodiments, sequencing reads are generated from a number of samples. In some embodiments, adaptors are ligated to a plurality of nucleic acid fragments from the samples, in which the nucleic acid fragments from each sample has the same indexing site. In some embodiments, a plurality of nucleic acid fragments is generated from a first sample and a second sample and adaptors are ligated to each nucleic acid fragment, in which adaptors ligated to each nucleic acid fragment from the first sample have the same first indexing site and adaptors ligated to nucleic acid fragments from the second sample has the same second indexing site. In some embodiments, the data associated with the nucleic acid fragments (e.g., sequencing reads) are separated based on the indexing site before sequencing reads of the target sequence and/or identifier site are analyzed. In some embodiments, the nucleic acid fragments or data associated with the nucleic acid fragments (e.g., sequencing reads) are separated based on the indexing site before duplicate sequencing reads are analyzed and/or removed.

[0070] In some embodiments, a method disclosed herein identifies or detects one or more true duplicate(s) with increased accuracy as compared to other methods. For example, in some embodiments, a method disclosed herein identify true duplicates (in contrast to identifying apparent or perceived duplicates) with increased accuracy as compared to other methods. The increased resolution and/or accuracy in identifying one or more true duplicate(s) can provide a considerable contribution to the state of the art in more accurately identifying true duplicates. In some embodiments, a method disclosed herein identifies or detects a true duplicate with

increased efficiency as compared to other methods, such as paired end sequencing. The increase in accuracy, resolution, and/or efficiency in detecting duplicate reads (e.g., true duplicate(s)) can increase confidence in sequencing results, such as for expression and CNV analysis.

Kits

[0071] Any of the compositions described herein may be comprised in a kit. In a non-limiting example, the kit, in a suitable container, comprises: an adaptor or several adaptors, one or more of oligonucleotide primers and reagents for amplification.

[0072] The containers of the kits will generally include at least one vial, test tube, flask, bottle, syringe or other containers, into which a component may be placed, and preferably, suitably aliquotted. Where there is more than one component in the kit, the kit also will generally contain a second, third or other additional container into which the additional components may be separately placed. However, various combinations of components may be comprised in a container.

[0073] When the components of the kit are provided in one or more liquid solutions, the liquid solution can be an aqueous solution. However, the components of the kit may be provided as dried powder(s). When reagents and/or components are provided as a dry powder, the powder can be reconstituted by the addition of a suitable solvent.

[0074] A kit may include instructions for employing the kit components as well the use of any other reagent not included in the kit. Instructions may include variations that can be implemented.

[0075] In some embodiments, the invention provides kits containing any one or more of the elements disclosed in the above methods and compositions. In some embodiments, a kit comprises a composition of the invention, in one or more containers. In some embodiments, the

invention provides kits comprising adapters, primers, and/or other oligonucleotides described herein. In some embodiments, the kit further comprises one or more of: (a) a DNA ligase, (b) a DNA-dependent DNA polymerase, (c) an RNA-dependent DNA polymerase, (d) a forward adapter (e) one or more oligonucleotides comprising reverse adaptor sequence and (f) one or more buffers suitable for one or more of the elements contained in said kit. The adapters, primers, other oligonucleotides, and reagents can be, without limitation, any of those described above. Elements of the kit can further be provided, without limitation, in any of the amounts and/or combinations (such as in the same kit or same container) described above. The kits may further comprise additional agents, such as those described above, for use according to the methods of the invention. For example, the kit can comprise a first forward adaptor that is a partial duplex adaptor as described herein, a second forward adapter, and a nucleic acid modifying enzyme specific for a restriction and/or cleavage site present in the first forward adaptor. The kit elements can be provided in any suitable container, including but not limited to test tubes, vials, flasks, bottles, ampules, syringes, or the like. The agents can be provided in a form that may be directly used in the methods of the invention, or in a form that requires preparation prior to use, such as in the reconstitution of lyophilized agents. Agents may be provided in aliquots for single-use or as stocks from which multiple uses, such as in a number of reaction, may be obtained.

[0076] In some embodiments, the kit comprises a plurality of adaptor oligonucleotides, wherein each of the adaptor oligonucleotides comprises at least one of a plurality of identifier site sequences, wherein each identifier site sequence of the plurality of identifier site sequences differs from every other identifier site sequence in said plurality of identifier site sequences at at least three nucleotide positions, and instructions for using the same. Adapters comprising

different identifier site sequences can be supplied individually or in combination with one or more additional adapters having a different identifier site sequence. In some embodiments, the kit can comprises a plurality of adapter oligonucleotides.

EXAMPLES

Example 1: Identification of Duplicate Sequencing Reads with NuGEN Ovation Target Enrichment Library System

[0077] Sample Description: l00ng of DNA from a human HapMap sample (NA19238) was fragmented to approximately 500 base pair in length by sonication with a Covaris system (Covaris, Inc., Woburn, MA). The resulting DNA was treated with end repair enzyme mix NuGEN R01280 and R01439 (NuGEN Technologies, Inc., San Carlos, CA) according to supplier's recommendation to produce blunt ended DNA fragments.

[0078] Library Generation, Enrichment, and Identifier Site Incorporation: An oligonucleotide with segments, from 5' to 3' of the top strand, 1) an Illumina indexing read priming site such as AGAGCACACGTCTGAACTCCAGTCAC (SEQ ID NO:2), 2) an indexing site, 3) an identifier site having a random 6 base sequence and, 4) a sequence compatible with an Illumina forward sequencing priming site such as TCTTTCCCTACACGACGCTCTTCCGATCT (SEQ ID NO:3) , was annealed to a second oligonucleotide to form a partially double stranded DNA adapter. Five uM of these adapters were ligated onto the end-repaired DNA using Ligase and Ligase reaction buffer from NuGEN's Ovation Ultralow Library System (NuGEN Technologies, Inc., San Carlos, CA) according to supplier's recommendations. Following 30 minutes of incubation at 25°C, the reaction mixture was diluted with water, 0.8X volume of Ampure XP magnetic beads (Agencourt Biosciences Corporation, A Beckman Coulter Company, Beverly, MA) was added and the solution thoroughly mixed. The beads were collected, washed and the ligated

DNA fragments eluted according to manufacturer's recommendations. A pool of targeting probes was annealed to the eluted DNA fragments by initially heating the solution to 95°C then slowly cooling the mixture from 80°C to 60°C by 0.6 degrees/minute. Targeting probes that were specifically annealed were extended with Taq DNA polymerase (New England Biolabs, Inc., Ipswich, MA) according to the manufacturer's protocols. Following extension, the DNA fragments were collected on Agencourt magnetic beads, washed and eluted according to manufacturer's recommendations. These libraries were enriched by 30 cycles of PCR using NuGEN library enrichment primers (Ovation Target Enrichment Library System, NuGEN Technologies, Inc., San Carlos, CA) that also contain the Illumina flow cell sequences (Illumina Inc., San Diego, CA) according to supplier's recommendation.

[0079] The resulting libraries were quantitated by qPCR using a kit provided by KAPA, diluted to 2nM and applied to an Illumina MiSeq DNA Sequencer (Illumina Inc., San Diego, CA). The following series was run: 36 base first read, 14 base second read, and 24 base third read.

[0080] *Data Analysis:* The sequencer output was processed in accordance with manufacturer's recommendation. In order to analyze the data, the indexing read was split into two files. The first file contained the first 8 bases of the indexing reads and is utilized as the library index file for standard library parsing. The other file contains only the random bases and is set aside for further sequence parsing.

[0081] Following our data analysis pipeline of sequence alignment with bowtie aligner (Langmead B. et al., Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* 2009, 10:R2.), duplicate reads were identified by their genomic start positions. At this point, sequencing reads that started at the same genomic position were checked against the random bases file to see if they have the same or different set of random

bases that were ligated to them. Where two sequences with the same starting genomic coordinate had the same set of random bases, they were considered to have come from the same initial DNA ligation event, regardless of which Ovation Target Enrichment targeting probe was used to generate the sequence fragment in question. These sequences, therefore, did not provide unique information about the starting genomic DNA and are considered as one sequencing read for the purpose of variant analysis. Two sequence reads with the same starting genomic coordinate that have different random bases were derived from unique ligation events and were considered to both be valid sequencing reads for the purpose of variant identification. **Figure 5** provides the analysis results demonstrating the identification of the duplicate reads. The use of an identifier site allowed for the determination of the number of true duplications versus the number of apparent or perceived duplicates.

[0082] If the sequences of two libraries are identical, their duplication status is unknown since this could occur by chance in any library. If an identifier sequence is used in combination with the library reads, the status can be determined (duplicates if identical, distinct if different). With the SPET system, one end is common so the probability of two libraries having identical ends increases. These would appear to be duplicate sequences and their true status can be determined by looking at an identifier sequence.

[0083] Over the sampling of all randomly selected reads, the use of identifier sites provided an increased resolution on the presence of true duplicates. When evaluating two million random reads, the apparent duplicates were found to comprise 39% of all reads. However, the true duplicates, identified through the use of an identifier site, were found to comprise only 26% of all reads. The methods employing the use of an identifier site were found to considerably increase the resolution of the true number of duplicates within the pool of reads.

Example 2: Removal of Duplicate Sequencing Reads with 8 Base Identifier Site

[0084] In a standard RNA sequencing library, adaptors are ligated to the ends of double stranded cDNA. These adaptors contain universal sequences that allow for PCR amplification and sequencing on high throughput sequencing machines. The adaptors are synthesized with a large population of additional sequences, in which each additional sequence is an identifier site, at the ligation end. The identifier sites are present at the junction between the adaptor and the cDNA. The sequence read starts with the identifier site and follows with the cDNA sequence.

[0085] This pool of identifier sites is used for the detection of PCR duplicates, as PCR duplicates will contain the identical identifier site, whereas two different cDNA molecules will ligate to two different adaptors containing two different identifier sites. This identifier site is designed as eight random bases introduced onto the end of the adaptor. Sequence reads from libraries made with such adaptors contain the 8 bases of the identifier site followed by the cDNA sequence. Standard PCR duplicate removal software such as PICARD, Markduplicates, and/or SAMtools rndup is used to identify and remove the PCR duplicates, leaving behind for analysis any instances of multiple cDNA fragments that happen to have the same sequence.

Example 3: Removal of Duplicate Sequencing Reads with Mixture of Random 1-8 Base Identifier Sites

[0086] In a standard RNA sequencing library, adaptors are ligated to the ends of double stranded cDNA. These adaptors contain universal sequences that allow for PCR amplification and sequencing on high throughput sequencing machines. The adaptors are synthesized with a large population of additional sequences, in which each additional sequence is an identifier site, at the

ligation end. The identifier sites are present at the junction between the adaptor and the cDNA. The sequence read starts with the identifier site and follows with the cDNA sequence.

[0087] This pool of identifier sites is used for the detection of PCR duplicates, as PCR duplicates will contain the identical identifier site, whereas two different cDNA molecules will ligate to two different adaptors containing two different identifier sites.

[0088] One to eight random bases are introduced onto the end of the adaptor. Sequence reads from libraries made with such adaptors contain between 1 and 8 bases of the identifier site followed by the cDNA sequence. Standard PCR duplicate removal software such as PICARD, Markduplicates, and/or SAMtools rndup is used to identify and remove the PCR duplicates, leaving behind for analysis any instances of multiple cDNA fragments that happen to have the same sequence.

Example 4: Removal of Duplicate Sequencing Reads with Mixture of 96 Defined 6-Base Identifier Sites

[0089] In a standard RNA sequencing library, adaptors are ligated to the ends of double stranded cDNA. These adaptors contain universal sequences that allow for PCR amplification and sequencing on high throughput sequencing machines. The adaptors are synthesized with a large population of additional sequences, in which each additional sequence is an identifier site, at the ligation end. The identifier sites are present at the junction between the adaptor and the cDNA. The sequence read starts with the identifier site and follows with the cDNA sequence.

[0090] This pool of identifier sites is used for the detection of PCR duplicates, as PCR duplicates will contain the identical identifier site, whereas two different cDNA molecules will ligate to two different adaptors containing two different identifier sites.

[0091] A mixture of 96 defined six-bases sequences are introduced onto the end of the adaptors. Thus, each six-base sequence is an identifier site. Sequence reads from libraries made with such adaptors contain one of the 96 six-base identifier site followed by the cDNA sequence. Standard PCR duplicate removal software such as PICARD, Markduplicates, and/or SAMtools rndup is used to identify and remove the PCR duplicates, leaving behind for analysis any instances of multiple cDNA fragments that happen to have the same sequence.

Example 5: Identification of Duplicate Sequencing Reads in Determining mRNA Expression Levels

[0092] Sample Description: Total RNA is extracted from tumor and normal adjacent tissue for the purpose of finding differences in expression levels of transcripts between the two sample types. 100ng of each sample is converted into cDNA using the USP primers, reaction buffer, and Reverse Transcriptase provided in NuGEN's Encore Complete Library System (NuGEN Technologies, Inc., San Carlos, CA) according to the supplier's recommendations. This is followed by second strand synthesis, again using materials provided in the kit according to recommendations. Double stranded cDNA was prepared using the SuperScript® Double-Stranded cDNA Synthesis Kit from Life Technologies (Carlsbad, CA) according the manufacturer's instructions. DNA was sheared with a Covaris S-series device (Covaris, Inc., Woburn, MA) using the 200 bp sonication protocol provided with the instrument (10% duty cycle, 200 cycles/burst, 5 intensity, 180 seconds). DNA was treated with 1.5 uL l0X Blunting Buffer, 0.5 uL Blunting Enzyme (New England Biolabs, Inc., Ipswich, MA; p/n E1201) and 1.2 uL of 2.5mM of each dNTP mix in a total volume of 15 uL for 30 minutes at 25°C followed by 10 minutes at 70°C.

[0093] *Library Generation*: The DNA fragments were then subjected to end repair using end repair buffers and enzymes provided in NuGEN's Ovation Ultralow Library System (NuGEN Technologies, Inc., San Carlos, CA).

[0094] Forward adaptor 5'

[0095] AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCC GATCTNNNNNNNN (SEQ ID NO:4),

[0096] Reverse adaptor 1) 5'

[0097] CAAGCAGAAGACGGCATACGAGATTCCCTTGTGACTGGAGTTCAGACGTGTG CTCTTCCGATCTNNNNNNNN (SEQ ID NO:5),

[0098] Reverse adaptor 2) 5'

[0099] CAAGCAGAAGACGGCATACGAGATTGAAGGGTGACTGGAGTTCAGACGTGT GCTCTTCCGATCNNNNNNNN (SEQ ID NO:6), and

[00100] Common partner 5' NNNNNNNNAGATCGGAAGAGC (SEQ ID NO:7) were all ordered from IDT (Integrated DNA Technologies, Coralville, IA). The reverse adaptors each contain a unique identifier (underlined) that enables libraries made with these adapters to be distinguished. In this case N represents an equimolar mixture of A, C, G, and T. A mixture of 5uM forward, 5uM reverse, and 10uM common in 10mM MgCl2, 50mM Tris pH 8 was heated to 95C for 5 minutes, then cooled to 20C. Adaptor ligation was performed by addition of 4.5 uL water, 3 uL Adaptor mix (prepared above), 6 uL 5X NEBNext Quick Ligation Reaction Buffer and 1.5 uL Quick T4 DNA Ligase (New England Biolabs, Inc., Ipswich, MA; p/n E6056), followed by incubation for 30 minutes at 25°C followed by 10 minutes at 70°C. Ligation products were purified by adding 70uL water and 80 uL of Ampure XP beads (Agencourt Genomics), washing twice with 70% ethanol and eluting with 20 uL of 10mM Tris pH 8.0.

Library products were amplified in a 50uL PCR containing 0.5uM each of primer (5' AATGATACGGCGACCACCGA (SEQ ID NO:8) , and 5' CAAGCAGAAGACGGCATACGA (SEQ ID NO:9), 10 mM Tris-HCl, pH 8.3, 50 mM KCl, 2 mM MgC12, 0.2mM each dNTP, and 1 unit Taq polymerase. The reaction was cycled 15 times under the conditions 95C for 15 seconds, 60C for 1 minute. PCR products were purified with 1 volume of Ampure XP beads (Agencourt Biosciences Corporation, A Beckman Coulter Company, Beverly, MA) as described above. The library was analyzed by HS DNA Bioanalyzer (Agilent Technologies, Santa Clara, CA) and quantitated with the KAPA Library Quantification Kit (KAPA Biosystems, Wilmington, MA; p/n KK4835) according to the supplied instructions. The resulting libraries are combined and are compatible with standard TruSeq single end or paired Illumina sequencing protocols for GAIIx, MiSeq, or Hi Seq sequencing instruments (Illumina Inc., San Diego, CA). The following series is run; 50 base first read, 6 base second read. A third read is not required for counting purposes or duplication analysis.

[00101] Data Analysis: The sequencer output was processed in accordance with manufacturer's recommendation. The 6 bases of the index read is used for standard library parsing, separating the data files from the two sample types. The first 50 bases of the target sequence read are compared to each other. Any read that has identical sequence to any other read is identified as a duplicate and removed from the population, thus, only a single copy is retained within the file. Once the duplicate reads have been removed, 8 bases are trimmed from each read. The trimmed reads are aligned to a reference genome. Differential expression is then determined by comparing FPKM (fragments per kilobase per million reads) values between libraries utilizing scripts such as cufflinks or cuffdiff (Trapnell et al. 2010, Nature Biotechnology, 28, 511-515; Trapnell et al. 2013, Nature Biotechnology, 31, 46-53).

Example 6: Sequencing of the Identifier Site With the Target Sequence in Paired End Reads

[00102] The Ovation Library System for Low Complexity Samples (NuGEN Technologies, Inc., San Carlos, CA) was used to generate four libraries, each from a single amplicon, following manufacturer's protocol. Purified libraries were mixed and sequenced as a multiplex on the Illumina MiSeq (Illumina Inc., San Diego, CA) to produce 125nt forward, 8nt index 1, 8nt index 2, and 25nt reverse reads. Because all amplicon reads start and end at the same sequence coordinates, the traditional method of marking library PCR duplicates (marking reads that start and end at the same genomic coordinates as duplicates) cannot be used. Instead, the 0-8nt of random sequence contained in the adaptors ligated to the amplicon were treated as an identifier sequence and used to mark duplicates. Any paired end reads that shared the same length and sequence of these random bases with any other paired end read was called a duplicate. The table below shows the results of this duplicate marking.

Table 1.

Library	Reads	Duplicate Reads	Reads from independent molecules
1	454249	376797	77452
2	439367	364001	75366
3	317760	253057	64703
4	476572	398380	78192

[00103] Table 1 demonstrates the accuracy of the method used to differentiate between duplicate reads and reads from truly independent molecules. The population of reads from independent molecules, depicted in the final column of Table 1, represent sequences from independent amplicon molecules used in generating the libraries.

Example 7: Sequencing of the Identifier Site With the Target Sequence in Reduced Representation Bisulfite (RRBS) Libraries

[00104] Reduced representation bisulfite (RRBS) libraries of the human genome are generated through the complete restriction enzyme digestion of 100ng input sample, followed by selection for short fragments. The resulting pool of fragments are ligated to adaptor sequences comprising an indexing site and an identifier site. The identifier sites comprise either 6 or 8 random nucleotides. The sequences are then sequenced to identify the identifier sites, thus revealing the true number of duplicates in the pool. In the absence of identifier sites, the number of apparent or perceived duplicates are greater than the number of true duplicates. The inclusion of an identifier site results in the identification of the number of true duplicates, as compared to the larger number of apparent or perceived duplicates.

[00105] Unless defined otherwise, all technical and scientific terms herein have the same meaning as commonly understood by one of ordinary skill in the art to which this invention belongs. Although any methods and materials, similar or equivalent to those described herein, can be used in the practice or testing of the present invention, the preferred methods and materials are described herein. All publications, patents, and patent publications cited are incorporated by reference herein in their entirety for all purposes.

[00106] The publications discussed herein are provided solely for their disclosure prior to the filing date of the present application. Nothing herein is to be construed as an admission that the present invention is not entitled to antedate such publication by virtue of prior invention.

[00107] All references, articles, publications, patents, patent publications, and patent applications cited herein are incorporated by reference in their entireties for all purposes. However, mention of any reference, article, publication, patent, patent publication, and

patent application cited herein is not, and should not be taken as an acknowledgment or any form of suggestion that they constitute valid prior art or form part of the common general knowledge in any country in the world.

[00108] While the invention has been described in connection with specific embodiments thereof, it will be understood that it is capable of further modifications and this application is intended to cover any variations, uses, or adaptations of the invention following, in general, the principles of the invention and including such departures from the present disclosure as come within known or customary practice within the art to which the invention pertains and as may be applied to the essential features hereinbefore set forth and as follows in the scope of the appended claims.

ABSTRACT

The present invention provides methods, compositions and kits for detecting duplicate sequencing reads. In some embodiments, the duplicate sequencing reads are removed.

CERTIFICATION AND REQUEST FOR PRIORITIZED EXAMINATION UNDER 37 CFR 1.102(e) (Page 1 of 1)

First Named Inventor:	AMORESE, Douglas	Nonprovisional Application Number (if known):	
Title of Invention:	COMPOSITIONS AND METHODS FOR IDENTIFICATION OF A DUPLICATE SEQUENCING READ		

APPLICANT HEREBY CERTIFIES THE FOLLOWING AND REQUESTS PRIORITIZED EXAMINATION FOR THE ABOVE-IDENTIFIED APPLICATION.

1. The processing fee set forth in 37 CFR 1.17(i)(1) and the prioritized examination fee set forth in 37 CFR 1.17(c) have been filed with the request. The publication fee requirement is met because that fee, set forth in 37 CFR 1.18(d), is currently $0. The basic filing fee, search fee, and examination fee are filed with the request or have been already been paid. I understand that any required excess claims fees or application size fee must be paid for the application.

2. I understand that the application may not contain, or be amended to contain, more than four independent claims, more than thirty total claims, or any multiple dependent claims, and that any request for an extension of time will cause an outstanding Track I request to be dismissed.

3. The applicable box is checked below:

I. ☒ Original Application (Track One) - Prioritized Examination under § 1.102(e)(1)

i. (a) The application is an original nonprovisional utility application filed under 35 U.S.C. 111(a). This certification and request is being filed with the utility application via EFS-Web.

---OR---

(b) The application is an original nonprovisional plant application filed under 35 U.S.C. 111(a). This certification and request is being filed with the plant application in paper.

ii. An executed inventor's oath or declaration under 37 CFR 1.63 or 37 CFR 1.64 for each inventor, **or** the application data sheet meeting the conditions specified in 37 CFR 1.53(f)(3)(i) is filed with the application.

II. ☐ Request for Continued Examination - Prioritized Examination under § 1.102(e)(2)

i. A request for continued examination has been filed with, or prior to, this form.
ii. If the application is a utility application, this certification and request is being filed via EFS-Web.
iii. The application is an original nonprovisional utility application filed under 35 U.S.C. 111(a), or is a national stage entry under 35 U.S.C. 371.
iv. This certification and request is being filed prior to the mailing of a first Office action responsive to the request for continued examination.
v. No prior request for continued examination has been granted prioritized examination status under 37 CFR 1.102(e)(2).

Signature /Angie Mah/	Date November 13, 2014
Name (Print/Typed) Angie S. Mah	Practitioner Registration Number 64,902

***Note**: This form must be signed in accordance with 37 CFR 1.33. See 37 CFR 1.4(d) for signature requirements and certifications. Submit multiple forms if more than one signature is required.**

☒ *Total of 1 forms are submitted.

American LegalNet, Inc.
www.FormsWorkFlow.com

Electronic Patent Application Fee Transmittal

Application Number:	
Filing Date:	
Title of Invention:	COMPOSITIONS AND METHODS FOR IDENTIFICATION OF A DUPLICATE SEQUENCING READ
First Named Inventor/Applicant Name:	Douglas AMORESE
Filer:	Angie S.Y. Mah
Attorney Docket Number:	NUGN-001/01US

Filed as Large Entity

Track I Prioritized Examination - Nonprovisional Application under 35 USC 111(a) Filing Fees

Description	Fee Code	Quantity	Amount	Sub-Total in USD($)
Basic Filing:				
Utility application filing	1011	1	280	280
Utility Search Fee	1111	1	600	600
Utility Examination Fee	1311	1	720	720
Request for Prioritized Examination	1817	1	4000	4000
Pages:				
Claims:				
Claims in Excess of 20	1202	1	80	80
Miscellaneous-Filing:				

Description	Fee Code	Quantity	Amount	Sub-Total in USD($)
Publ. Fee- Early, Voluntary, or Normal	1504	1	0	0
PROCESSING FEE, EXCEPT PROV. APPLS.	1830	1	140	140
Petition:				
Patent-Appeals-and-Interference:				
Post-Allowance-and-Post-Issuance:				
Extension-of-Time:				
Miscellaneous:				
			Total in USD ($)	**5820**

CLAIMS

1. A method for detecting a duplicate sequencing read from a population of sample sequencing reads comprising:
    a) ligating an adaptor to a 5' end of each nucleic acid fragment of a plurality of nucleic acid fragments from one or more samples, wherein the adaptor comprises:
        (i) an indexing primer binding site;
        (ii) an indexing site;
        (iii) an identifier site; and
        (iv) a target sequence primer binding site;
    b) amplifying the adapter-nucleic acid fragment ligated products;
    c) generating a population of sequencing reads from amplified adapter-nucleic acid fragment ligated products; and
    d) detecting the population of sequencing reads comprising a sequencing read with a duplicate identifier site and target sequence.
2. The method of claim 1, wherein the method further comprises removing from the population of sequence reads the sequencing read with a duplicate identifier site and target sequence.
3. The method of claim 1, wherein the identifier site is sequenced with the indexing site.
4. The method of claim 1, wherein the identifier site is sequenced separately from the indexing site.
5. The method of claim 1, wherein the identifier site is sequenced with the target sequence.
6. The method of claim 1, wherein the identifier site is sequenced separately from the target sequence.
7. The method of claim 1, wherein the adaptor comprises from 5' to 3':

(i) the indexing primer binding site;

(ii) the indexing site;

(iii) the identifier site; and

(iv) the target sequence primer binding site.

8. The method of claim 1, wherein the adaptor comprises from 5' to 3':

(i) the indexing primer binding site;

(ii) the indexing site;

(iii) the target sequence primer binding site; and

(iv) the identifier site.

9. The method of claim 1, wherein the plurality of nucleic acid fragments is generated from more than one sample.

10. The method of claim 9, wherein the nucleic acid fragments from each sample has the same indexing site.

11. The method of claim 10, wherein the sequencing reads are separated based on the indexing site.

12. The method of claim 11, wherein the separation of sequencing reads is performed prior to step d).

13. The method of claim 1, wherein the nucleic acid fragments are DNA fragments, RNA fragments, or DNA/RNA fragments.

14. The method of claim 13, wherein the nucleic acid fragments are genomic DNA fragments or cDNA fragments.

15. The method of claim 1, wherein the indexing site is between 2 and 8 nucleotides in length.

16. The method of claim 1, wherein the indexing site is about 6 nucleotides in length.

17. The method of claim 1, wherein the identifier site is between 1 and 8 nucleotides in length.

18. The method of claim 1, wherein the identifier site is about 8 nucleotides in length.

19. The method of claim 1, wherein the indexing primer binding site is a universal indexing primer binding site.

20. The method of claim 1, wherein the target sequence primer binding site is a universal target sequence primer binding site.

21. A kit comprising a plurality of adaptors, wherein each adaptor comprises:

    (i) an indexing primer binding site;

    (ii) an indexing site;

    (iii) an identifier site; and

    (iv) a target sequencing primer binding site.

PTO/AIA/15 (03-13)
Approved for use through 01/31/2014. 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.

UTILITY PATENT APPLICATION TRANSMITTAL

(Only for new nonprovisional applications under 37 CFR 1.53(b))

Attorney Docket No.	NUGN-001/01US
First Named Inventor	Doug AMORESE
Title	COMPOSITIONS AND METHODS FOR IDENTIFICATION OF A DUPLICATE SEQUE
Express Mail Label No.	

APPLICATION ELEMENTS

See MPEP chapter 600 concerning utility patent application contents.

1. [ ] **Fee Transmittal Form** (PTO/SB/17 or equivalent)
2. [ ] **Applicant asserts small entity status.** See 37 CFR 1.27
3. [ ] **Applicant certifies micro entity status.** See 37 CFR 1.29. Applicant must attach form PTO/SB/15A or B or equivalent.
4. [✓] **Specification** [*Total Pages* 44] Both the claims and abstract must start on a new page. *(See MPEP § 608.01(a) for information on the preferred arrangement)*
5. [✓] **Drawing(s)** (35 U.S.C. 113) [*Total Sheets* 6]
6. **Inventor's Oath or Declaration** [*Total Pages* ____] *(including substitute statements under 37 CFR 1.64 and assignments serving as an oath or declaration under 37 CFR 1.63(e))*
   a. [ ] Newly executed (original or copy)
   b. [ ] A copy from a prior application (37 CFR 1.63(d))
7. [✓] **Application Data Sheet** ** See note below.* See 37 CFR 1.76 (PTO/AIA/14 or equivalent)
8. **CD-ROM or CD-R** in duplicate, large table, or Computer Program (*Appendix*)
   [ ] Landscape Table on CD
9. **Nucleotide and/or Amino Acid Sequence Submission** (*if applicable, items a. – c. are required*)
   a. [✓] Computer Readable Form (CRF)
   b. [ ] Specification Sequence Listing on:
      i. [ ] CD-ROM or CD-R (2 copies); or
      ii. [ ] Paper
   c. [ ] Statements verifying identity of above copies

ADDRESS TO:

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

ACCOMPANYING APPLICATION PAPERS

10. [ ] **Assignment Papers** (cover sheet & document(s))
    Name of Assignee _____
11. [ ] **37 CFR 3.73(c) Statement** (*when there is an assignee*) [ ] **Power of Attorney**
12. [ ] **English Translation Document** (*if applicable*)
13. [ ] **Information Disclosure Statement** (PTO/SB/08 or PTO-1449)
    [ ] Copies of citations attached
14. [ ] **Preliminary Amendment**
15. [ ] **Return Receipt Postcard** (*MPEP § 503*) (*Should be specifically itemized*)
16. [ ] **Certified Copy of Priority Document(s)** (*if foreign priority is claimed*)
17. [ ] **Nonpublication Request** Under 35 U.S.C. 122(b)(2)(B)(i). Applicant must attach form PTO/SB/35 or equivalent.
18. [✓] **Other:** Certification and Request for Prioritized Examination Under 37 CFR 1.102(e)

Note:* (1) Benefit claims under 37 CFR 1.78 and foreign priority claims under 1.55 **must be included in an Application Data Sheet (ADS).
(2) For applications filed under 35 U.S.C. 111, the application must contain an ADS specifying the applicant if the applicant is an assignee, person to whom the inventor is under an obligation to assign, or person who otherwise shows sufficient proprietary interest in the matter. See 37 CFR 1.46(b).

19. CORRESPONDENCE ADDRESS

[✓] The address associated with Customer Number: 58249 **OR** [ ] Correspondence address below

Name					
Address					
City		State		Zip Code	
Country		Telephone		Email	

Signature	/Angie Mah/	Date	Nov. 13, 2014
Name (Print/Type)	Angie Mah	Registration No. (Attorney/Agent)	64902

This collection of information is required by 37 CFR 1.53(b). 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 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.

DocCode – SCORE

SCORE Placeholder Sheet for IFW Content

Application Number: 14540917 Document Date: 11/13/2014

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 eDAN using the Supplemental Content tab, or via the SCORE web page.
- External customers may access SCORE content via PAIR using the Supplemental Content tab.

Form Revision Date: August 26, 2013

DocCode – SEQ.TXT

SCORE Placeholder Sheet for IFW Content

Application Number: 14540917 Document Date: 11/13/2014

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.

- Sequence Listing

At the time of document entry (noted above):

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

Form Revision Date: August 26, 2013

Document code: WFEE

United States Patent and Trademark Office
Sales Receipt for Accounting Date: 11/25/2014

ABALINAN SALE #00000018 Mailroom Dt: 11/13/2014 501283 14540917
01 FC : 1051 140.00 DA

==========
Sequence Listing was accepted.
If you need help call the Patent Electronic Business Center at (866) 217-9197 (toll free).
Reviewer: Anjum, Durreshwar (CGI Federal)
Timestamp: [year=2014; month=11; day=14; hr=9; min=57; sec=36; ms=55;]
==========

Validated By CRFValidator v 1.0.4

Application No: 14540917 Version No: 1.1

Input Set:

Output Set:

Started: 2014-11-14 09:55:49.808
Finished: 2014-11-14 09:55:51.963
Elapsed: 0 hr(s) 0 min(s) 2 sec(s) 155 ms
Total Warnings: 47
Total Errors: 0
No. of SeqIDs Defined: 9
Actual SeqID Count: 9

Error code		Error Description
W	213	Artificial or Unknown found in <213> in SEQ ID (1)
W	447	n or Xaa used, for: SEQID(1) on line number 38
W	447	n or Xaa used, for: SEQID(1) on line number 38
W	447	n or Xaa used, for: SEQID(1) on line number 38
W	447	n or Xaa used, for: SEQID(1) on line number 38
W	447	n or Xaa used, for: SEQID(1) on line number 38
W	447	n or Xaa used, for: SEQID(1) on line number 38
W	213	Artificial or Unknown found in <213> in SEQ ID (2)
W	213	Artificial or Unknown found in <213> in SEQ ID (3)
W	213	Artificial or Unknown found in <213> in SEQ ID (4)
W	447	n or Xaa used, for: SEQID(4) on line number 82
W	447	n or Xaa used, for: SEQID(4) on line number 82
W	447	n or Xaa used, for: SEQID(4) on line number 84
W	447	n or Xaa used, for: SEQID(4) on line number 84
W	447	n or Xaa used, for: SEQID(4) on line number 84
W	447	n or Xaa used, for: SEQID(4) on line number 84
W	447	n or Xaa used, for: SEQID(4) on line number 84
W	447	n or Xaa used, for: SEQID(4) on line number 84
W	213	Artificial or Unknown found in <213> in SEQ ID (5)
W	447	n or Xaa used, for: SEQID(5) on line number 104

Input Set:

Output Set:

Started: 2014-11-14 09:55:49.808
Finished: 2014-11-14 09:55:51.963
Elapsed: 0 hr(s) 0 min(s) 2 sec(s) 155 ms
Total Warnings: 47
Total Errors: 0
No. of SeqIDs Defined: 9
Actual SeqID Count: 9

Error code	Error Description
W 447	n or Xaa used, for: SEQID(5) on line number 104
W 447	n or Xaa used, for: SEQID(5) on line number 104
W 447	n or Xaa used, for: SEQID(5) on line number 104
W 447	n or Xaa used, for: SEQID(5) on line number 104
W 447	n or Xaa used, for: SEQID(5) on line number 104 This error has occured more than 20 times, will not be displayed
W 213	Artificial or Unknown found in <213> in SEQ ID (6)
W 213	Artificial or Unknown found in <213> in SEQ ID (7)
W 213	Artificial or Unknown found in <213> in SEQ ID (8)
W 213	Artificial or Unknown found in <213> in SEQ ID (9)

SEQUENCE LISTING

<110> NUGEN TECHNOLOGIES, INC.
AMORESE, DOUGLAS
SCOLNICK, JONATHAN
SCHROEDER, BEN

<120> COMPOSITIONS AND METHODS FOR IDENTIFICATION
OF A DUPLICATE SEQUENCING READ

<130> NUGN-001/01US

<140> US 14/540,917
<141> 2014-11-13

<150> US 61/903,826
<151> 2013-11-13

<160> 9

<170> PatentIn version 3.5

<210> 1
<211> 70
<212> DNA
<213> Artificial Sequence

<220>
<223> Putative adaptor sequence

<220>
<221> misc_feature
<222> (34)..(39)
<223> n is a, c, g, or t

<400> 1
cagagcacac gtctgaactc cagtcacagt gagnnnnnna cactttccct acacgacgct 60

cttccgatct 70

<210> 2
<211> 26
<212> DNA
<213> Artificial Sequence

<220>
<223> Indexing read priming site

<400> 2
agagcacacg tctgaactcc agtcac 26

<210> 3
<211> 29
<212> DNA

<213> Artificial Sequence

<220>
<223> Illumina forward sequencing priming site

<400> 3
tctttcccta cacgacgctc ttccgatct 29

<210> 4
<211> 66
<212> DNA
<213> Artificial Sequence

<220>
<223> Forward adaptor

<220>
<221> misc_feature
<222> (59)..(66)
<223> n is a, c, g, or t

<400> 4
aatgatacgg cgaccaccga gatctacact ctttccctac acgacgctct tccgatctnn 60

nnnnnn 66

<210> 5
<211> 72
<212> DNA
<213> Artificial Sequence

<220>
<223> Reverse adaptor 1

<220>
<221> misc_feature
<222> (65)..(72)
<223> n is a, c, g, or t

<400> 5
caagcagaag acggcatacg agattccctt gtgactggag ttcagacgtg tgctcttccg 60

atctnnnnnn nn 72

<210> 6
<211> 71
<212> DNA
<213> Artificial Sequence

<220>
<223> Reverse adaptor 2

<220>
<221> misc_feature
<222> (64)..(71)
<223> n is a, c, g, or t

<400> 6
caagcagaag acggcatacg agattgaagg gtgactggag ttcagacgtg tgctcttccg 60

atcnnnnnnn n 71

<210> 7
<211> 21
<212> DNA
<213> Artificial Sequence

<220>
<223> Common partner

<220>
<221> misc_feature
<222> (1)..(8)
<223> n is a, c, g, or t

<400> 7
nnnnnnnnag atcggaagag c 21

<210> 8
<211> 20
<212> DNA
<213> Artificial Sequence

<220>
<223> Library amplification primer 1

<400> 8
aatgatacgg cgaccaccga 20

<210> 9
<211> 21
<212> DNA
<213> Artificial Sequence

<220>
<223> Library amplification primer 2

<400> 9
caagcagaag acggcatacg a 21

PTO/SB/08a (01-10)
Approved for use through 07/31/2012. OMB 0651-0031
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 contains a valid OMB control number.

INFORMATION DISCLOSURE STATEMENT BY APPLICANT

(Not for submission under 37 CFR 1.99)

Application Number	14540917
Filing Date	2014-11-13
First Named Inventor	Douglas AMORESE
Art Unit	1636
Examiner Name	Unassigned
Attorney Docket Number	NUGN-001/01US

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	6251639	B1	2001-06-26	Nurith Kurn	
	2	5750341	A	1998-05-12	Stephen C. Macevicz	
	3	6306597	B1	2001-10-23	Stephen C. Macevicz	
	4	5969119	A	1999-10-19	Stephen C. Macevicz	
	5	7244559	B2	2007-07-17	Rothberg et al.	
	6	7335762	B2	2008-02-26	Rothberg et al.	
	7	7211390	B2	2007-05-01	Rothberg et al.	
	8	7244567	B2	2007-07-17	Chen et al.	

INFORMATION DISCLOSURE STATEMENT BY APPLICANT (Not for submission under 37 CFR 1.99)		
	Application Number	14540917
	Filing Date	2014-11-13
	First Named Inventor	Douglas AMORESE
	Art Unit	1636
	Examiner Name	Unassigned
	Attorney Docket Number	NUGN-001/01US

	9	7264929	B2	2007-09-04	Rothberg et al.	
	10	7323305	B2	2008-01-29	Leamon et al.	
	11	7501245	B2	2009-03-10	Quake et al.	
	12	7491498	B2	2009-02-17	Lapidus et al.	
	13	7276720	B2	2007-10-02	Kevin Ulmer	
	14	7462452	B2	2008-12-09	Williams et al.	
	15	7476504	B2	2009-01-13	Stephen Turner	
	16	7405281	B2	2008-07-29	Xu et al.	
	17	7170050	B2	2007-01-30	Turner et al.	
	18	7462468	B1	2008-12-09	Williams et al.	
	19	7476503	B2	2009-01-13	Turner et al.	

INFORMATION DISCLOSURE STATEMENT BY APPLICANT (Not for submission under 37 CFR 1.99)		
	Application Number	14540917
	Filing Date	2014-11-13
	First Named Inventor	Douglas AMORESE
	Art Unit	1636
	Examiner Name	Unassigned
	Attorney Docket Number	NUGN-001/01US

	20	7315019	B2	2008-01-01	Turner et al.	
	21	7302146	B2	2007-11-27	Turner et al.	
	22	7313318	B2	2007-12-25	Turner et al.	
	23	6449562	B1	2002-10-10	Chandler et al.	
	24	6287766	B1	2001-09-11	Nolan et al.	
	25	7361468	B2	2008-04-22	Liu et al.	
	26	7414117	B2	2008-08-19	Saito et al.	
	27	6225109	B1	1999-05-01	Juncosa et al.	
	28	6110709	A	2000-08-29	Ausubel et al.	

If you wish to add additional U.S. Patent citation information please click the Add button.

U.S.PATENT APPLICATION PUBLICATIONS

Examiner Initial*	Cite No	Publication Number	Kind Code1	Publication Date	Name of Patentee or Applicant of cited Document	Pages,Columns,Lines where Relevant Passages or Relevant Figures Appear

INFORMATION DISCLOSURE STATEMENT BY APPLICANT (Not for submission under 37 CFR 1.99)		
	Application Number	14540917
	Filing Date	2014-11-13
	First Named Inventor	Douglas AMORESE
	Art Unit	1636
	Examiner Name	Unassigned
	Attorney Docket Number	NUGN-001/01US

	1	20130231253	A1	2013-09-05	Amorese et al.	
	2	20060024711	A1	2006-02-02	Lapidus et al.	
	3	20090061439	B2	2010-08-03	Philip Richard Buzby	
	4	20080087826	A1	2012-11-01	Gore et al.	
	5	20060286566	A1	2006-12-21	Lapidus et al.	
	6	20060024711	A1	2006-02-02	Lapidus et al.	
	7	20060024678	A1	2006-02-02	Philip Buzby	
	8	20080213770	A1	2008-09-04	Williams et al.	
	9	20080103058	A1	2008-05-01	Suhaib Siddiqi	
	10	20090029385	A1	2009-01-29	Christians et al.	
	11	20090068655	A1	2009-03-12	John G.K. Williams	

INFORMATION DISCLOSURE STATEMENT BY APPLICANT (Not for submission under 37 CFR 1.99)		
	Application Number	14540917
	Filing Date	2014-11-13
	First Named Inventor	Douglas AMORESE
	Art Unit	1636
	Examiner Name	Unassigned
	Attorney Docket Number	NUGN-001/01US

	12	20090024331	A1	2009-01-22	Tomaney et al.	
	13	20080206764	A1	2008-18-28	Williams et al.	
	14	20090026082	A1	2009-01-29	Rothberg et al.	

If you wish to add additional U.S. Published Application citation information please click the Add button.

FOREIGN PATENT DOCUMENTS

Examiner Initial*	Cite No	Foreign Document Number[3]	Country Code[2] i	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]
	1							☐

If you wish to add additional Foreign Patent Document citation information please click the Add button

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, pages(s), volume-issue number(s), publisher, city and/or country where published.	T[5]
	1	WESTIN et al., Anchored multiplex amplification on a microelectronic chip array. 2000. Nature Biotechnology. 18:199-204.	☐
	2	Walker et al., Strand displacement amplification--an isothermal, in vitro DNA amplification technique. 1992. Nucleic Acids Research. 20(7):1691-1696.	☐
	3	MARGULIES et al. Genome sequencing in open microfabricated high density picoliter reactors. 2005. Nature. 437 (7057):376-380.	☐

IDT Exhibit 1018.0077

INFORMATION DISCLOSURE STATEMENT BY APPLICANT (Not for submission under 37 CFR 1.99)		
	Application Number	14540917
	Filing Date	2014-11-13
	First Named Inventor	Douglas AMORESE
	Art Unit	1636
	Examiner Name	Unassigned
	Attorney Docket Number	NUGN-001/01US

	4	SONI and MELLER. Progress toward ultrafast DNA sequencing using solid-state nanopores. 2007. Clinical Chemistry. 53(11):1996-2001.	☐
	5	LANGMEAD et al. Ultragast and memory-efficient alignment of short DNA sequences to the human genome. 2009. Genome Biology. 10:R25.	☐
	6	TRAPNELL et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. 2010. Nature Biotechnology. 28:511-515.	☐
	7	TRAPNELL et al. Differential analysis of gene regulation at transcript resolution with RNA-seq. 2013. Nature Biotechnology. 31:46-53.	☐

If you wish to add additional non-patent literature document citation information please click the Add button

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.

INFORMATION DISCLOSURE STATEMENT BY APPLICANT

(Not for submission under 37 CFR 1.99)

Application Number	14540917
Filing Date	2014-11-13
First Named Inventor	Douglas AMORESE
Art Unit	1636
Examiner Name	Unassigned
Attorney Docket Number	NUGN-001/01US

CERTIFICATION STATEMENT

Please see 37 CFR 1.97 and 1.98 to make the appropriate selection(s):

☐ That each item of information contained in the 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 the information disclosure statement. See 37 CFR 1.97(e)(1).

OR

☐ That no item of information contained in the information disclosure statement was cited in a communication from a foreign patent office in a counterpart foreign application, and, to the knowledge of the person signing the certification after making reasonable inquiry, no item of information contained in the information disclosure statement was known to any individual designated in 37 CFR 1.56(c) more than three months prior to the filing of the information disclosure statement. See 37 CFR 1.97(e)(2).

☐ See attached certification statement.

☐ The fee set forth in 37 CFR 1.17 (p) has been submitted herewith.

☒ A certification statement is not submitted herewith.

SIGNATURE

A signature of the applicant or representative is required in accordance with CFR 1.33, 10.18. Please see CFR 1.4(d) for the form of the signature.

Signature	/Angie Mah/	Date (YYYY-MM-DD)	2014-12-04
Name/Print	Angie Mah	Registration Number	64902

This collection of information is required by 37 CFR 1.97 and 1.98. 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 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: Commissioner for Patents, P.O. Box 1450, Alexandria, VA 22313-1450.**

IDT Exhibit 1018.0079

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 the Freedom of Information Act requires disclosure of these record s.

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 inspections 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.

IDT Exhibit 1018.0080

Electronic Acknowledgement Receipt

EFS ID:	20863062
Application Number:	14540917
International Application Number:	
Confirmation Number:	7076
Title of Invention:	COMPOSITIONS AND METHODS FOR IDENTIFICATION OF A DUPLICATE SEQUENCING READ
First Named Inventor/Applicant Name:	Douglas AMORESE
Customer Number:	58249
Filer:	Angie S.Y. Mah
Filer Authorized By:	
Attorney Docket Number:	NUGN-001/01US
Receipt Date:	04-DEC-2014
Filing Date:	13-NOV-2014
Time Stamp:	14:09:11
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	Information Disclosure Statement (IDS) Form (SB08)	NUGN-001_01US_IDS_Dec_2014.pdf	75555 a4a75fba5f684ee5e9dafae6d2f5baecb9f9f389	no	8

Warnings:	
Information:	

This is not an USPTO supplied IDS fillable form					
2	Non Patent Literature	Langmead_2009.pdf	418882 617ee080bc27157a4689294d8e6de1f177062853	no	10
Warnings:					
Information:					
3	Non Patent Literature	Margulies_2005.pdf	746855 e43ecea5bb529f9e0058f2e961bce09f03575081	no	14
Warnings:					
Information:					
4	Non Patent Literature	Soni_and_Meller_2007.pdf	383298 93a337c85efb504915007effed5a4797944231d1	no	6
Warnings:					
Information:					
5	Non Patent Literature	Trapnell_2010.pdf	1463680 b7884f48141ee94f82108b312ac8a59d72d557e5	no	18
Warnings:					
Information:					
6	Non Patent Literature	Trapnell_2013.pdf	1564250 0bd974367512f5ee9e44aca4578d10e3d84aa884	no	19
Warnings:					
Information:					
7	Non Patent Literature	Walker_1992.pdf	11569484 f4b9a2c1656ed35008682dd74468ce2910230609	no	6
Warnings:					
Information:					
8	Non Patent Literature	Westin_2000.pdf	298035 941b7f18f1bc8946e11c60c988efb41422ed9f1b	no	6
Warnings:					
Information:					
			Total Files Size (in bytes):	16520039	

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 NUMBER	FILING OR 371(C) DATE	FIRST NAMED APPLICANT	ATTY. DOCKET NO./TITLE
14/540,917	11/13/2014	Douglas AMORESE	NUGN-001/01US

CONFIRMATION NO. 7076

58249
COOLEY LLP
ATTN: Patent Group
1299 Pennsylvania Avenue, NW
Suite 700
Washington, DC 20004

NOTICE

OC000000072136865

Date Mailed: 12/04/2014

INFORMATIONAL NOTICE TO APPLICANT

Applicant is notified that the above-identified application contains the deficiencies noted below. No period for reply is set forth in this notice for correction of these deficiencies. However, if a deficiency relates to the inventor's oath or declaration, the applicant must file an oath or declaration in compliance with 37 CFR 1.63, or a substitute statement in compliance with 37 CFR 1.64, executed by or with respect to each actual inventor no later than the expiration of the time period set in the "Notice of Allowability" to avoid abandonment. See 37 CFR 1.53(f).

The item(s) indicated below are also required and should be submitted with any reply to this notice to avoid further processing delays.

- A properly executed inventor's oath or declaration has not been received for the following inventor(s):
  Douglas AMORESE
  Jonathan Scolnick
  Ben Schroeder

IDT Exhibit 1018.0084

PATENT APPLICATION FEE DETERMINATION RECORD

Substitute for Form PTO-875

Application or Docket Number: 14/540,917

APPLICATION AS FILED - PART I

FOR	(Column 1) NUMBER FILED	(Column 2) NUMBER EXTRA	SMALL ENTITY RATE($)	SMALL ENTITY FEE($)	OR	OTHER THAN SMALL ENTITY RATE($)	OTHER THAN SMALL ENTITY FEE($)
BASIC FEE (37 CFR 1.16(a), (b), or (c))	N/A	N/A	N/A			N/A	280
SEARCH FEE (37 CFR 1.16(k), (i), or (m))	N/A	N/A	N/A			N/A	600
EXAMINATION FEE (37 CFR 1.16(o), (p), or (q))	N/A	N/A	N/A			N/A	720
TOTAL CLAIMS (37 CFR 1.16(i))	21 minus 20 =	* 1			OR	x 80 =	80
INDEPENDENT CLAIMS (37 CFR 1.16(h))	2 minus 3 =	*				x 420 =	0.00
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).						0.00
MULTIPLE DEPENDENT CLAIM PRESENT (37 CFR 1.16(j))							0.00
* If the difference in column 1 is less than zero, enter "0" in column 2.			TOTAL			TOTAL	1680

APPLICATION AS AMENDED - PART II

		(Column 1) CLAIMS REMAINING AFTER AMENDMENT		(Column 2) HIGHEST NUMBER PREVIOUSLY PAID FOR	(Column 3) PRESENT EXTRA	SMALL ENTITY RATE($)	SMALL ENTITY ADDITIONAL FEE($)	OR	OTHER THAN SMALL ENTITY RATE($)	OTHER THAN SMALL ENTITY ADDITIONAL FEE($)
AMENDMENT A	Total (37 CFR 1.16(i))	*	Minus	**	=	x =		OR	x =	
	Independent (37 CFR 1.16(h))	*	Minus	***	=	x =		OR	x =	
	Application Size Fee (37 CFR 1.16(s))									
	FIRST PRESENTATION OF MULTIPLE DEPENDENT CLAIM (37 CFR 1.16(j))							OR		
						TOTAL ADD'L FEE		OR	TOTAL ADD'L FEE	

		(Column 1) CLAIMS REMAINING AFTER AMENDMENT		(Column 2) HIGHEST NUMBER PREVIOUSLY PAID FOR	(Column 3) PRESENT EXTRA	SMALL ENTITY RATE($)	SMALL ENTITY ADDITIONAL FEE($)	OR	OTHER THAN SMALL ENTITY RATE($)	OTHER THAN SMALL ENTITY ADDITIONAL FEE($)
AMENDMENT B	Total (37 CFR 1.16(i))	*	Minus	**	=	x =		OR	x =	
	Independent (37 CFR 1.16(h))	*	Minus	***	=	x =		OR	x =	
	Application Size Fee (37 CFR 1.16(s))									
	FIRST PRESENTATION OF MULTIPLE DEPENDENT CLAIM (37 CFR 1.16(j))							OR		
						TOTAL ADD'L FEE		OR	TOTAL ADD'L FEE	

* If the entry in column 1 is less than the entry in column 2, write "0" in column 3.

** If the "Highest Number Previously Paid For" IN THIS SPACE is less than 20, enter "20".

*** 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 found in the appropriate box in column 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

APPLICATION NUMBER	FILING or 371(c) DATE	GRP ART UNIT	FIL FEE REC'D	ATTY.DOCKET.NO	TOT CLAIMS	IND CLAIMS
14/540,917	11/13/2014	1636	1820	NUGN-001/01US	21	2

CONFIRMATION NO. 7076

58249
COOLEY LLP
ATTN: Patent Group
1299 Pennsylvania Avenue, NW
Suite 700
Washington, DC 20004

FILING RECEIPT

OC000000072136864

Date Mailed: 12/04/2014

Receipt is acknowledged of this non-provisional patent application. The application will be taken up for examination in due course. Applicant will be notified as to the results of the examination. Any correspondence concerning the application must include the following identification information: the U.S. APPLICATION NUMBER, FILING DATE, NAME OF APPLICANT, and TITLE OF INVENTION. Fees transmitted by check or draft are subject to collection. Please verify the accuracy of the data presented on this receipt. **If an error is noted on this Filing Receipt, please submit a written request for a Filing Receipt Correction. Please provide a copy of this Filing Receipt with the changes noted thereon. If you received a "Notice to File Missing Parts" for this application, please submit any corrections to this Filing Receipt with your reply to the Notice. When the USPTO processes the reply to the Notice, the USPTO will generate another Filing Receipt incorporating the requested corrections**

Inventor(s)

Douglas AMORESE, Los Altos, CA;
Jonathan Scolnick, San Francisco, CA;
Ben Schroeder, San Mateo, CA;

Applicant(s)

Nugen Technologies, Inc., San Carlos, CA

Power of Attorney: None

Domestic Priority data as claimed by applicant

This appln claims benefit of 61/903,826 11/13/2013

Foreign Applications for which priority is claimed (You may be eligible to benefit from the **Patent Prosecution Highway** program at the USPTO. Please see http://www.uspto.gov for more information.) - None.
Foreign application information must be provided in an Application Data Sheet in order to constitute a claim to foreign priority. See 37 CFR 1.55 and 1.76.

Permission to Access - A proper **Authorization to Permit Access to Application by Participating Offices** (PTO/SB/39 or its equivalent) has been received by the USPTO.

If Required, Foreign Filing License Granted: 11/25/2014
The country code and number of your priority application, to be used for filing abroad under the Paris Convention, is **US 14/540,917**
Projected Publication Date: 05/14/2015
Non-Publication Request: No
Early Publication Request: No

IDT Exhibit 1018.0086

Title

COMPOSITIONS AND METHODS FOR IDENTIFICATION OF A DUPLICATE SEQUENCING READ

Preliminary Class

435

Statement under 37 CFR 1.55 or 1.78 for AIA (First Inventor to File) Transition Applications: No

PROTECTING YOUR INVENTION OUTSIDE THE UNITED STATES

Since the rights granted by a U.S. patent extend only throughout the territory of the United States and have no effect in a foreign country, an inventor who wishes patent protection in another country must apply for a patent in a specific country or in regional patent offices. Applicants may wish to consider the filing of an international application under the Patent Cooperation Treaty (PCT). An international (PCT) application generally has the same effect as a regular national patent application in each PCT-member country. The PCT process **simplifies** the filing of patent applications on the same invention in member countries, but **does not result** in a grant of "an international patent" and does not eliminate the need of applicants to file additional documents and fees in countries where patent protection is desired.

Almost every country has its own patent law, and a person desiring a patent in a particular country must make an application for patent in that country in accordance with its particular laws. Since the laws of many countries differ in various respects from the patent law of the United States, applicants are advised to seek guidance from specific foreign countries to ensure that patent rights are not lost prematurely.

Applicants also are advised that in the case of inventions made in the United States, the Director of the USPTO must issue a license before applicants can apply for a patent in a foreign country. The filing of a U.S. patent application serves as a request for a foreign filing license. The application's filing receipt contains further information and guidance as to the status of applicant's license for foreign filing.

Applicants may wish to consult the USPTO booklet, "General Information Concerning Patents" (specifically, the section entitled "Treaties and Foreign Patents") for more information on timeframes and deadlines for filing foreign patent applications. The guide is available either by contacting the USPTO Contact Center at 800-786-9199, or it can be viewed on the USPTO website at http://www.uspto.gov/web/offices/pac/doc/general/index.html.

For information on preventing theft of your intellectual property (patents, trademarks and copyrights), you may wish to consult the U.S. Government website, http://www.stopfakes.gov. Part of a Department of Commerce initiative, this website includes self-help "toolkits" giving innovators guidance on how to protect intellectual property in specific countries such as China, Korea and Mexico. For questions regarding patent enforcement issues, applicants may call the U.S. Government hotline at 1-866-999-HALT (1-866-999-4258).

IDT Exhibit 1018.0087

LICENSE FOR FOREIGN FILING UNDER

Title 35, United States Code, Section 184

Title 37, Code of Federal Regulations, 5.11 & 5.15

GRANTED

The applicant has been granted a license under 35 U.S.C. 184, if the phrase "IF REQUIRED, FOREIGN FILING LICENSE GRANTED" followed by a date appears on this form. Such licenses are issued in all applications where the conditions for issuance of a license have been met, regardless of whether or not a license may be required as set forth in 37 CFR 5.15. The scope and limitations of this license are set forth in 37 CFR 5.15(a) unless an earlier license has been issued under 37 CFR 5.15(b). The license is subject to revocation upon written notification. The date indicated is the effective date of the license, unless an earlier license of similar scope has been granted under 37 CFR 5.13 or 5.14.

This license is to be retained by the licensee and may be used at any time on or after the effective date thereof unless it is revoked. This license is automatically transferred to any related applications(s) filed under 37 CFR 1.53(d). This license is not retroactive.

The grant of a license does not in any way lessen the responsibility of a licensee for the security of the subject matter as imposed by any Government contract or the provisions of existing laws relating to espionage and the national security or the export of technical data. Licensees should apprise themselves of current regulations especially with respect to certain countries, of other agencies, particularly the Office of Defense Trade Controls, Department of State (with respect to Arms, Munitions and Implements of War (22 CFR 121-128)); the Bureau of Industry and Security, Department of Commerce (15 CFR parts 730-774); the Office of Foreign AssetsControl, Department of Treasury (31 CFR Parts 500+) and the Department of Energy.

NOT GRANTED

No license under 35 U.S.C. 184 has been granted at this time, if the phrase "IF REQUIRED, FOREIGN FILING LICENSE GRANTED" DOES NOT appear on this form. Applicant may still petition for a license under 37 CFR 5.12, if a license is desired before the expiration of 6 months from the filing date of the application. If 6 months has lapsed from the filing date of this application and the licensee has not received any indication of a secrecy order under 35 U.S.C. 181, the licensee may foreign file the application pursuant to 37 CFR 5.15(b).

SelectUSA

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 U.S. offers tremendous resources and advantages for those who invest and manufacture goods here. Through SelectUSA, our nation works to promote and facilitate business investment. SelectUSA provides information assistance to the international investor community; serves as an ombudsman for existing and potential investors; advocates on behalf of U.S. cities, states, and regions competing for global investment; and counsels U.S. economic development organizations on investment attraction best practices. To learn more about why the United States is the best country in the world to develop technology, manufacture products, deliver services, and grow your business, visit http://www.SelectUSA.gov or call +1-202-482-6800.

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

MAILED
DEC 10 2014
OFFICE OF PETITIONS

COOLEY LLP
ATTN: PATENT GROUP
1299 PENNSYLVANIA AVENUE, NW
SUITE 700
WASHINGTON DC 20004

Doc Code: TRACK1.GRANT

Decision Granting Request for Prioritized Examination (Track I or After RCE)	Application No.: 14/540,917

1. THE REQUEST FILED November 13, 2014 IS **GRANTED**.

The above-identified application has met the requirements for prioritized examination
- A. ☒ for an original nonprovisional application (Track I).
- B. ☐ for an application undergoing continued examination (RCE).

2. **The above-identified application will undergo prioritized examination.** The application will be accorded special status throughout its entire course of prosecution until one of the following occurs:

- A. filing a **petition for extension of time** to extend the time period for filing a reply;
- B. filing an **amendment to amend the application to contain more than four independent claims, more than thirty total claims**, or a multiple dependent claim;
- C. filing a **request for continued examination**;
- D. filing a notice of appeal;
- E. filing a request for suspension of action;
- F. mailing of a notice of allowance;
- G. mailing of a final Office action;
- H. completion of examination as defined in 37 CFR 41.102; or
- I. abandonment of the application.

Telephone inquiries with regard to this decision should be directed to **Michelle R. Eason** at (571) 272-4231.

In his/her absence, calls may be directed to Brian W. Brown at (571) 272-5338.

/Michelle R. Eason/
(*Signature*)

Paralegal Specialist, Office of Petitions
(Title)

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

In Re Application of:	AMORESE, Douglas, et al.	Confirmation No.:	7076
Serial No.:	14/540,917	Group Art Unit:	1636
Filed:	November 13, 2014	Examiner:	*To be assigned*

FOR: **COMPOSITIONS AND METHODS FOR IDENTIFICATION OF A DUPLICATE SEQUENCING READ**

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

POWER BY ASSIGNEE AND STATEMENT UNDER 37 C.F.R. §3.73(c)

The Assignee of the entire right, title, and interest in the above-identified application hereby grants the registered practitioners of Cooley LLP included in the Customer Number provided below power to act, prosecute, and transact all business in the U.S. Patent and Trademark Office in connection with this application, any applications claiming priority to this application, and any patents issuing therefrom.

The Assignee certifies that to the best of its knowledge and belief it is the owner of the entire right, title, and interest in and to the above-identified application as evidenced by:

☒ An assignment document, a copy of which is enclosed herewith and is being recorded with the U.S. Patent and Trademark Office concurrently herewith;

☐ An assignment previously recorded in the U.S. Patent and Trademark Office at Reel ___, Frame ___.

Please direct all telephone calls and correspondence to:

CUSTOMER NUMBER: **58249**

COOLEY LLP
ATTN: Patent Group
1299 Pennsylvania Avenue NW, Suite 700
Washington, DC 20004
Tel: (202) 842-7800
Fax: (202) 842-7899

The undersigned (whose title is supplied below) is empowered to sign this statement on behalf of the Assignee.

Date: 11 DEC 14

Signature: _____

Name: ELIZABETH A. HUNT

Title: CEO

Company: NuGen Technologies, Inc.

ASSIGNMENT

AMORESE, Douglas, residing at Los Altos, California, and **SCHROEDER, Ben**, residing at San Mateo, California, and **SCOLNICK, Jonathan**, residing at San Francisco, California, (each referred to as "Assignor") have made an invention(s) (the "Invention(s)") set forth in an application for patent of the United States, entitled **COMPOSITIONS AND METHODS FOR IDENTIFICATION OF A DUPLICATE SEQUENCING READ**, and which is a:

(1) ☐ provisional application
- (a) ☐ to be filed herewith; or
- (b) ☐ bearing Application No. , and filed on ;

(2) ☒ non-provisional application
- (a) ☐ to be filed herewith; or
- (b) ☒ bearing Application No. 14/540,917, and filed on November 13, 2014; and/or

(3) ☐ PCT application
- (a) ☐ bearing Application No. , and filed on .

WHEREAS, NuGen Technologies, Inc., a corporation duly organized under and pursuant to the laws of California, and having its principal place of business at 201 Industrial Road, Suite 310, San Carlos, California 94070 (the "Assignee"), is desirous of acquiring the entire right, title, and interest in: the Invention(s); the application(s) for patent identified in paragraph (1), (2) and/or (3); the right to file applications for patent of the United States or other countries on the Invention(s); any application(s) for patent of the United States or other countries claiming priority to, and/or the benefit of, these applications; any provisional or other right to recover damages, including royalties, for prior infringements of these application(s); and any patent(s) of the United States or other countries that may be granted therefor or thereon.

NOW, THEREFORE, for good and sufficient consideration, the receipt of which is hereby acknowledged, and to the extent that the Assignor has not done so already via a prior agreement with the Assignee, or if the Assignor has already done so via a prior agreement with the Assignee then in confirmation of any obligation to do so in said prior agreement, the Assignor has sold, assigned, transferred, and set over, and by these presents does sell, assign, transfer, and set over, unto the Assignee, its successors, legal representatives, and assigns, the Assignor's entire right, title, and interest in:

(a) the Invention(s);

(b) the application(s) for patent identified in paragraph (1), (2) and/or (3);

(c) the right to file applications for patent of the United States or other countries on the Invention(s), including all rights under the Hague Convention, the Paris Convention for the Protection of Industrial Property, and under the Patent Cooperation Treaty, and all rights of claiming priority in any country of the world;

(d) any application(s) for patent of the United States or other countries claiming the Invention(s);

(e) any application(s) for patent of the United States or other countries claiming priority to, and/or the benefit of, at least one of the application(s) for patent identified in paragraph (1), (2) and/or (3) or any application(s) for patent claiming the Invention(s), including any substitute application(s), division(s), continuation(s), and continuation(s)-in-part;

(f) any provisional or other right to recover damages, including royalties, for prior infringements of any application for patent identified in the preceding paragraphs (b)-(e); and

(g) any patent(s) of the United States or other countries that may be granted for or on any application for patent identified in the preceding paragraphs (b)-(e), including any reissue(s), reexamination(s), revival(s), renewal(s) and extension(s) of said patent(s).

The above-granted rights, titles, and interests are to be held and enjoyed by the Assignee, for its own use and behalf and the use and behalf of its successors, legal representatives, and assigns, as fully and entirely as the same would have been held and enjoyed by the Assignor had this sale and assignment not been made.

The Assignor hereby represents to the Assignee, its successors, legal representatives, and assigns, that, at the time of execution and delivery of these presents, or if applicable, at such time said prior agreement was executed, the Assignor is a lawful owner of an undivided interest in the entire right, title, and interest in and to the Invention(s), that the Invention(s) are unencumbered, except, if applicable, by obligation to assign in accordance with said prior agreement, and that the Assignor has good and full right and lawful authority to sell and convey the same in the manner set forth herein, and that Assignor will not make or enter into any assignment, sale, agreement or encumbrance which would conflict with these presents.

The Assignor hereby covenants and agrees to and with the Assignee, its successors, legal representatives, and assigns, that the Assignor will sign all papers and documents, take all lawful oaths, and do all acts necessary or required to be done in connection with any and all proceedings for the procurement, maintenance, enforcement and defense of the Invention(s), said application(s), and said patent(s), including interference and derivation proceedings, and any post-grant proceedings (e.g., opposition

IDT Exhibit 1018.0093

proceedings, post-grant reviews, *Inter partes* reviews, supplemental examinations, etc.) without charge to the Assignee, its successors, legal representatives, and assigns, but at the cost and expense of the Assignee, its successors, legal representatives, and assigns.

The Assignor hereby authorizes and requests the attorneys of COOLEY LLP to insert in the spaces provided above the filing date, the application number, and the attorney docket number of the application(s) identified in paragraph (1), (2) and/or (3) when known.

The Assignor hereby requests the Commissioner of Patents to issue said patents of the United States to the Assignee for the sole use and behalf of the Assignee, its successors, legal representatives, and assigns.

Date: Dec 11, 2014 By: Douglas Amorese

State of CALIFORNIA)
) ss.
County of SANTA CLARA)

On 12/11/14, before me, KYLE DEVIN EISENBERG, Notary Public, personally appeared DOUGLAS ALAN AMORESE, who proved to me on the basis of satisfactory evidence, to be the person(s) whose name(s) is/are subscribed to the within instrument and acknowledged to me that he/she/they executed the same in his/her/their authorized capacity(ies), and that by his/her/their signature(s) on the instrument the person(s), or the entity upon behalf of which the person(s) acted, executed the instrument.

REQUIRED SENTENCE IF NOTARIZED IN CALIFORNIA: I certify under PENALTY OF PERJURY under the laws of the State of California that the foregoing paragraph is true and correct.

KYLE DEVIN EISENBERG
Commission # 2067666
Notary Public - California
San Francisco County
My Comm. Expires May 10, 2018

WITNESS my hand and official seal.

Signature of Notary Public

Place Notary Seal Above

My Commission Expires: 5-10-2018

Date: 11 Dec 2014 By: [signature]
Ben Schroeder

State of CALIFORNIA)
) ss.
County of SANTA CLARA)

On 12/11/14, before me, KYLE DEVIN EISENBERG, Notary Public, personally appeared BENJAMIN GEORGE SCHROEDER, who proved to me on the basis of satisfactory evidence, to be the person(s) whose name(s) is/are subscribed to the within instrument and acknowledged to me that he/she/they executed the same in his/her/their authorized capacity(ies), and that by his/her/their signature(s) on the instrument the person(s), or the entity upon behalf of which the person(s) acted, executed the instrument.

REQUIRED SENTENCE IF NOTARIZED IN CALIFORNIA: I certify under PENALTY OF PERJURY under the laws of the State of California that the foregoing paragraph is true and correct.

WITNESS my hand and official seal.

KYLE DEVIN EISENBERG
Commission # 2067568
Notary Public - California
San Francisco County
My Comm. Expires May 10, 2018

[signature]
Signature of Notary Public

Place Notary Seal Above

My Commission Expires: 5-10-18

111916878 v2

Date: 11 December 14 By: _____
Jonathan Scolnick

State of CALIFORNIA)
) ss.
County of SANTA CLARA)

On 12/11/14, before me, KYLE DEVIN EISENBERG, Notary Public, personally appeared JONATHAN ADAM SCOLNICK, who proved to me on the basis of satisfactory evidence, to be the person(s) whose name(s) is/are subscribed to the within instrument and acknowledged to me that he/she/they executed the same in his/her/their authorized capacity(ies), and that by his/her/their signature(s) on the instrument the person(s), or the entity upon behalf of which the person(s) acted, executed the instrument.

REQUIRED SENTENCE IF NOTARIZED IN CALIFORNIA: I certify under PENALTY OF PERJURY under the laws of the State of California that the foregoing paragraph is true and correct.

WITNESS my hand and official seal.

KYLE DEVIN EISENBERG
Commission # 2067666
Notary Public - California
San Francisco County
My Comm. Expires May 10, 2018

Signature of Notary Public

Place Notary Seal Above

My Commission Expires: 5-10-18

Date: 11 DEC 14 By: _____
Name: ELIZABETH A. HUTT
Title: CEO
Company: NuGen Technologies, Inc.

State of CALIFORNIA)
) ss.
County of SANTA CLARA)

On 12/11/14, before me, KYLE DEVIN EISENBERG, Notary Public, personally appeared ELIZABETH ANN HUTT POLLARD, who proved to me on the basis of satisfactory evidence, to be the person(s) whose name(s) is/are subscribed to the within instrument and acknowledged to me that he/she/they executed the same in his/her/their authorized capacity(ies), and that by his/her/their signature(s) on the instrument the person(s), or the entity upon behalf of which the person(s) acted, executed the instrument.

REQUIRED SENTENCE IF NOTARIZED IN CALIFORNIA: I certify under PENALTY OF PERJURY under the laws of the State of California that the foregoing paragraph is true and correct.

KYLE DEVIN EISENBERG
Commission # 2067666
Notary Public - California
San Francisco County
My Comm. Expires May 10, 2018

WITNESS my hand and official seal.

Signature of Notary Public

Place Notary Seal Above

My Commission Expires: 5-10-2018

IDT Exhibit 1018.0098

Electronic Acknowledgement Receipt

EFS ID:	21077651
Application Number:	14540917
International Application Number:	
Confirmation Number:	7076
Title of Invention:	COMPOSITIONS AND METHODS FOR IDENTIFICATION OF A DUPLICATE SEQUENCING READ
First Named Inventor/Applicant Name:	Douglas AMORESE
Customer Number:	58249
Filer:	Angie S.Y. Mah
Filer Authorized By:	
Attorney Docket Number:	NUGN-001/01US
Receipt Date:	31-DEC-2014
Filing Date:	13-NOV-2014
Time Stamp:	08:15:14
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	Oath or Declaration filed	NUGN_001_01US_Decs.pdf	132009 c8cd06b8f7ad2226ccdfcdda6b1aac392f3d8ea9	no	3

Warnings:

The page size in the PDF is too large. The pages should be 8.5 x 11 or A4. If this PDF is submitted, the pages will be resized upon entry into the Image File Wrapper and may affect subsequent processing					
Information:					
2	Power of Attorney	NUGN_001_01US_POA.pdf	254191 1223823626f57d91c8794ef732aaf0ce9c219ff4	no	9
Warnings:					
The page size in the PDF is too large. The pages should be 8.5 x 11 or A4. If this PDF is submitted, the pages will be resized upon entry into the Image File Wrapper and may affect subsequent processing					
Information:					
			Total Files Size (in bytes):	386200	

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.

PTO/AIA/01 (06-12)
Approved for use through 01/31/2014. 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.

DECLARATION (37 CFR 1.63) FOR UTILITY OR DESIGN APPLICATION USING AN APPLICATION DATA SHEET (37 CFR 1.76)

Title of Invention	COMPOSITIONS AND METHODS FOR IDENTIFICATION OF A DUPLICATE SEQUENCING READ

As the below named inventor, I hereby declare that:

This declaration is directed to:

☐ The attached application, or

☒ United States application or PCT international application number 14/540,917

filed on 11/13/2014.

The above-identified application was made or authorized to be made by me.

I believe that I am the original inventor or an original joint inventor of a claimed invention in the application.

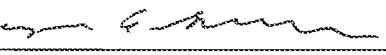
I hereby acknowledge that any willful false statement made in this declaration is punishable under 18 U.S.C. 1001 by fine or imprisonment of not more than five (5) years, or both.

WARNING:

Petitioner/applicant is cautioned to avoid submitting personal information in documents filed in a patent application that may contribute to identity theft. Personal information such as social security numbers, bank account numbers, or credit card numbers (other than a check or credit card authorization form PTO-2038 submitted for payment purposes) is never required by the USPTO to support a petition or an application. If this type of personal information is included in documents submitted to the USPTO, petitioners/applicants should consider redacting such personal information from the documents before submitting them to the USPTO. Petitioner/applicant is advised that the record of a patent application is available to the public after publication of the application (unless a non-publication request in compliance with 37 CFR 1.213(a) is made in the application) or issuance of a patent. Furthermore, the record from an abandoned application may also be available to the public if the application is referenced in a published application or an issued patent (see 37 CFR 1.14). Checks and credit card authorization forms PTO-2038 submitted for payment purposes are not retained in the application file and therefore are not publicly available.

LEGAL NAME OF INVENTOR

Inventor: Ben Schroeder Date (Optional) : _____

Signature: 

Note: An application data sheet (PTO/SB/14 or equivalent), including naming the entire inventive entity, must accompany this form or must have been previously filed. Use an additional PTO/AIA/01 form for each additional inventor.

This collection of information is required by 35 U.S.C. 115 and 37 CFR 1.63. 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 1 minute 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.

American LegalNet, Inc.
www.FormsWorkFlow.com

PTO/AIA/01 (06-12)
Approved for use through 01/31/2014. 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.

DECLARATION (37 CFR 1.63) FOR UTILITY OR DESIGN APPLICATION USING AN APPLICATION DATA SHEET (37 CFR 1.76)

Title of Invention	COMPOSITIONS AND METHODS FOR IDENTIFICATION OF A DUPLICATE SEQUENCING READ

As the below named inventor, I hereby declare that:

This declaration is directed to:

☐ The attached application, or

☒ United States application or PCT international application number 14/540,917

filed on 11/13/2014.

The above-identified application was made or authorized to be made by me.

I believe that I am the original inventor or an original joint inventor of a claimed invention in the application.

I hereby acknowledge that any willful false statement made in this declaration is punishable under 18 U.S.C. 1001 by fine or imprisonment of not more than five (5) years, or both.

WARNING:

Petitioner/applicant is cautioned to avoid submitting personal information in documents filed in a patent application that may contribute to identity theft. Personal information such as social security numbers, bank account numbers, or credit card numbers (other than a check or credit card authorization form PTO-2038 submitted for payment purposes) is never required by the USPTO to support a petition or an application. If this type of personal information is included in documents submitted to the USPTO, petitioners/applicants should consider redacting such personal information from the documents before submitting them to the USPTO. Petitioner/applicant is advised that the record of a patent application is available to the public after publication of the application (unless a non-publication request in compliance with 37 CFR 1.213(a) is made in the application) or issuance of a patent. Furthermore, the record from an abandoned application may also be available to the public if the application is referenced in a published application or an issued patent (see 37 CFR 1.14). Checks and credit card authorization forms PTO-2038 submitted for payment purposes are not retained in the application file and therefore are not publicly available.

LEGAL NAME OF INVENTOR

Inventor: Douglas Amorese Date (Optional) :

Signature:

Note: An application data sheet (PTO/SB/14 or equivalent), including naming the entire inventive entity, must accompany this form or must have been previously filed. Use an additional PTO/AIA/01 form for each additional inventor.

This collection of information is required by 35 U.S.C. 115 and 37 CFR 1.63. 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 1 minute 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.

American LegalNet, Inc.
www.FormsWorkFlow.com

PTO/AIA/01 (06-12)
Approved for use through 01/31/2014. 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.

DECLARATION (37 CFR 1.63) FOR UTILITY OR DESIGN APPLICATION USING AN APPLICATION DATA SHEET (37 CFR 1.76)

Title of Invention	COMPOSITIONS AND METHODS FOR IDENTIFICATION OF A DUPLICATE SEQUENCING READ

As the below named inventor, I hereby declare that:

This declaration is directed to:

☐ The attached application, or

☒ United States application or PCT international application number 14/540,917

filed on 11/13/2014.

The above-identified application was made or authorized to be made by me.

I believe that I am the original inventor or an original joint inventor of a claimed invention in the application.

I hereby acknowledge that any willful false statement made in this declaration is punishable under 18 U.S.C. 1001 by fine or imprisonment of not more than five (5) years, or both.

WARNING:

Petitioner/applicant is cautioned to avoid submitting personal information in documents filed in a patent application that may contribute to identity theft. Personal information such as social security numbers, bank account numbers, or credit card numbers (other than a check or credit card authorization form PTO-2038 submitted for payment purposes) is never required by the USPTO to support a petition or an application. If this type of personal information is included in documents submitted to the USPTO, petitioners/applicants should consider redacting such personal information from the documents before submitting them to the USPTO. Petitioner/applicant is advised that the record of a patent application is available to the public after publication of the application (unless a non-publication request in compliance with 37 CFR 1.213(a) is made in the application) or issuance of a patent. Furthermore, the record from an abandoned application may also be available to the public if the application is referenced in a published application or an issued patent (see 37 CFR 1.14). Checks and credit card authorization forms PTO-2038 submitted for payment purposes are not retained in the application file and therefore are not publicly available.

LEGAL NAME OF INVENTOR

Inventor: Jonathan Scolnick Date (Optional) : _____

Signature: _____

Note: An application data sheet (PTO/SB/14 or equivalent), including naming the entire inventive entity, must accompany this form or must have been previously filed. Use an additional PTO/AIA/01 form for each additional inventor.

This collection of information is required by 35 U.S.C. 115 and 37 CFR 1.63. 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 1 minute 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.

American LegalNet, Inc.
www.FormsWorkFlow.com

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.
14/540,917	11/13/2014	Douglas AMORESE	NUGN-001/01US	7076

58249 7590 01/07/2015

COOLEY LLP
ATTN: Patent Group
1299 Pennsylvania Avenue, NW
Suite 700
Washington, DC 20004

EXAMINER
STRZELECKA, TERESA E

ART UNIT	PAPER NUMBER
1637	

NOTIFICATION DATE	DELIVERY MODE
01/07/2015	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):

zpatdcdocketing@cooley.com

IDT Exhibit 1018.0104

Office Action Summary	Application No. 14/540,917	Applicant(s) AMORESE ET AL.	
	Examiner TERESA E. STRZELECKA	Art Unit 1637	AIA (First Inventor to File) Status Yes

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

Period for Reply

A SHORTENED STATUTORY PERIOD FOR REPLY IS SET TO EXPIRE 2 MONTHS FROM THE MAILING DATE OF THIS COMMUNICATION.

- Extensions of time may be available under the provisions of 37 CFR 1.136(a). In no event, however, may a reply be timely filed after SIX (6) MONTHS from the mailing date of this communication.
- If NO period for reply is specified above, the maximum statutory period will apply and will expire SIX (6) MONTHS from the mailing date of this communication.
- Failure to reply within the set or extended period for reply will, by statute, cause the application to become ABANDONED (35 U.S.C. § 133). Any reply received by the Office later than three months after the mailing date of this communication, even if timely filed, may reduce any earned patent term adjustment. See 37 CFR 1.704(b).

Status

1)☐ Responsive to communication(s) filed on _____.

☐ A declaration(s)/affidavit(s) under **37 CFR 1.130(b)** was/were filed on _____.

2a)☐ This action is **FINAL**. 2b)☐ This action is non-final.

3)☐ 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.

4)☐ Since this application is in condition for allowance except for formal matters, prosecution as to the merits is closed in accordance with the practice under *Ex parte Quayle*, 1935 C.D. 11, 453 O.G. 213.

Disposition of Claims*

5)☒ Claim(s) *1-21* is/are pending in the application.

5a) Of the above claim(s) _____ is/are withdrawn from consideration.

6)☐ Claim(s) _____ is/are allowed.

7)☐ Claim(s) _____ is/are rejected.

8)☐ Claim(s) _____ is/are objected to.

9)☒ Claim(s) *1-21* are subject to restriction and/or election requirement.

* If any claims have been determined allowable, 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.

Application Papers

10)☐ The specification is objected to by the Examiner.

11)☐ The drawing(s) filed on _____ is/are: a)☐ accepted or b)☐ objected to by the Examiner.

Applicant may not request that any objection to the drawing(s) be held in abeyance. See 37 CFR 1.85(a).

Replacement drawing sheet(s) including the correction is required if the drawing(s) is objected to. See 37 CFR 1.121(d).

Priority under 35 U.S.C. § 119

12)☐ 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)).

** See the attached detailed Office action for a list of the certified copies not received.

Attachment(s)

1) ☐ Notice of References Cited (PTO-892)

2) ☐ Information Disclosure Statement(s) (PTO/SB/08a and/or PTO/SB/08b) Paper No(s)/Mail Date _____.

3) ☐ Interview Summary (PTO-413) Paper No(s)/Mail Date. _____ .

4) ☐ Other: _____.

IDT Exhibit 1018.0105

1. The present application, filed on or after March 16, 2013, is being examined under the first inventor to file provisions of the AIA.

DETAILED ACTION

Election/Restrictions

2. Restriction to one of the following inventions is required under 35 U.S.C. 121:

I. Claims 1-20, drawn to a method for detecting a duplicate sequencing read from a population of sample sequencing reads comprising:

a) ligating an adaptor to a 5' end of each nucleic acid fragment of a plurality of nucleic acid fragments from one or more samples, wherein the adaptor comprises:

(i) an indexing primer binding site;

(ii) an indexing site;

(iii) an identifier site; and

(iv) a target sequence primer binding site;

b) amplifying the adapter-nucleic acid fragment ligated products;

c) generating a population of sequencing reads from amplified adapter-nucleic acid fragment ligated products; and

d) detecting the population of sequencing reads comprising a sequencing read with a duplicate identifier site and target sequence, classified in C12Q 1/6806, for example.

II. Claim 21, drawn to a kit comprising a plurality of adaptors, wherein each adaptor comprises:

(i) an indexing primer binding site;

(ii) an indexing site;

(iii) an identifier site; and

(iv) a target sequencing primer binding site, classified in C12Q 6855, for example.

The inventions are distinct, each from the other because of the following reasons:

3. Inventions I and II are related as product and process of use. The inventions can be shown to be distinct if either or both of the following can be shown: (1) the process for using the product as claimed can be practiced with another materially different product or (2) the product as claimed can be used in a materially different process of using that product. See MPEP § 806.05(h). In the instant case the kit of Group II can be used in an entirely different method, such as library construction from different tissues, rather than in the method of Group I.

Searching the inventions of Groups I and II together would impose serious search burden. The inventions of Groups I and II have a separate status in the art as shown by their different classifications. Moreover, in the instant case, the search for the adaptor oligonucleotide and the method of sequencing using adaptor oligonucleotides are not coextensive. Group II encompasses all possible oligonucleotides, since the terms "primer binding site", "indexing site", identifier site" and "target sequence binding site" are not claimed in terms of their specific structure, i.e., a nucleic acid sequence. In contrast, the search for group I would require a text search for the method of using adaptors in sequencing. Prior art which teaches an adaptor of Group II would not necessarily be applicable to the method of using adaptors in sequencing. Moreover, even if the adaptor product were known, the method of using the product may be novel and unobvious in view of the preamble or active steps.

4. Restriction for examination purposes as indicated is proper because all these inventions listed in this action are independent or distinct for the reasons given above and there would be a

serious search and/or examination burden if restriction were not required because one or more of the following reasons apply:

Because these inventions are distinct for the reasons given above, have acquired a separate status in the art as shown by their different classification, and the search required for each group is not required for the other groups because each group requires a different non-patent literature search due to each group comprising different products and/or method steps, restriction for examination purposes as indicated is proper.

Applicant is advised that the reply to this requirement to be complete must include (i) an election of an invention to be examined even though the requirement may be traversed (37 CFR 1.143) **and (ii) identification of the claims encompassing the elected invention**.

The election of an invention may be made with or without traverse. To reserve a right to petition, the election must be made with traverse. If the reply does not distinctly and specifically point out supposed errors in the restriction requirement, the election shall be treated as an election without traverse. Traversal must be presented at the time of election in order to be considered timely. Failure to timely traverse the requirement will result in the loss of right to petition under 37 CFR 1.144. If claims are added after the election, applicant must indicate which of these claims are readable upon the elected invention.

Should applicant traverse on the ground that the inventions are not patentably distinct, applicant should submit evidence or identify such evidence now of record showing the inventions to be obvious variants or clearly admit on the record that this is the case. In either instance, if the examiner finds one of the inventions unpatentable over the prior art, the evidence

or admission may be used in a rejection under 35 U.S.C. 103 or pre-AIA 35 U.S.C. 103(a) of the other invention.

5. This application contains claims directed to the following patentably distinct species:

Group I

Species of sequencing the identifier site

A) identifier site is sequenced with the indexing site (claim 3),

B) identifier site is sequenced separately from the indexing site (claim 4),

C) identifier site is sequenced with the target sequence (claim 5),

D) identifier site is sequenced separately from the target sequence (claim 6).

Species of adaptor

E) adaptor comprises from 5' to 3': (i) the indexing primer binding site; (ii) the indexing site; (iii) the identifier site; and (iv) the target sequence primer binding site (claim 7),

F) adaptor comprises from 5' to 3': (i) the indexing primer binding site; (ii) the indexing site; (iii) the target sequence primer binding site; and (iv) the identifier site (claim 8).

Species of nucleic acid fragments

G) genomic DNA fragments (claim 13, 14),

H) cDNA fragments (claim 13, 14),

I) RNA fragments (claim 13),

J) DNA/RNA fragments (claim 13).

6. The species are independent or distinct because claims to the different species recite the mutually exclusive characteristics of such species. In addition, these species are not obvious variants of each other based on the current record.

Applicant is required under 35 U.S.C. 121 to elect a single disclosed species from each of the sets of species if Group I is elected, or a single grouping of patentably indistinct species, for prosecution on the merits to which the claims shall be restricted if no generic claim is finally held to be allowable. Currently, claim 1 is generic.

There is a search and/or examination burden for the patentably distinct species as set forth above because at least the following reason(s) apply:

The species require a different field of search (e.g., searching different classes/subclasses or electronic resources, or employing different search queries); and/or the prior art applicable to one species would not likely be applicable to another species; and/or the species are likely to raise different non-prior art issues under 35 U.S.C. 101 and/or 35 U.S.C. 112, first paragraph.

Applicant is advised that the reply to this requirement to be complete must include (i) an election of a species to be examined even though the requirement may be traversed (37 CFR 1.143) **and (ii) identification of the claims encompassing the elected species or grouping of patentably indistinct species**, including any claims subsequently added. An argument that a claim is allowable or that all claims are generic is considered nonresponsive unless accompanied by an election.

The election may be made with or without traverse. To preserve a right to petition, the election must be made with traverse. If the reply does not distinctly and specifically point out supposed errors in the election of species requirement, the election shall be treated as an election without traverse. Traversal must be presented at the time of election in order to be considered timely. Failure to timely traverse the requirement will result in the loss of right to petition under

37 CFR 1.144. If claims are added after the election, applicant must indicate which of these claims are readable on the elected species or grouping of patentably indistinct species.

Should applicant traverse on the ground that the species, or groupings of patentably indistinct species from which election is required, are not patentably distinct, applicant should submit evidence or identify such evidence now of record showing them to be obvious variants or clearly admit on the record that this is the case. In either instance, if the examiner finds one of the species unpatentable over the prior art, the evidence or admission may be used in a rejection under 35 U.S.C. 103 or pre-AIA 35 U.S.C. 103(a) of the other species.

Upon the allowance of a generic claim, applicant will be entitled to consideration of claims to additional species which depend from or otherwise require all the limitations of an allowable generic claim as provided by 37 CFR 1.141.

7. Applicant is reminded that upon the cancellation of claims to a non-elected invention, the inventorship must be corrected in compliance with 37 CFR 1.48(a) if one or more of the currently named inventors is no longer an inventor of at least one claim remaining in the application. A request to correct inventorship under 37 CFR 1.48(a) must be accompanied by an application data sheet in accordance with 37 CFR 1.76 that identifies each inventor by his or her legal name and by the processing fee required under 37 CFR 1.17(i).

8. The examiner has required restriction between product or apparatus claims and process claims. Where applicant elects claims directed to the product/apparatus, and all product/apparatus claims are subsequently found allowable, withdrawn process claims that include all the limitations of the allowable product/apparatus claims should be considered for

rejoinder. All claims directed to a nonelected process invention must include all the limitations of an allowable product/apparatus claim for that process invention to be rejoined.

In the event of rejoinder, the requirement for restriction between the product/apparatus claims and the rejoined process claims will be withdrawn, and the rejoined process claims will be fully examined for patentability in accordance with 37 CFR 1.104. Thus, to be allowable, the rejoined claims must meet all criteria for patentability including the requirements of 35 U.S.C. 101, 102, 103 and 112. Until all claims to the elected product/apparatus are found allowable, an otherwise proper restriction requirement between product/apparatus claims and process claims may be maintained. Withdrawn process claims that are not commensurate in scope with an allowable product/apparatus claim will not be rejoined. See MPEP § 821.04. Additionally, in order for rejoinder to occur, applicant is advised that the process claims should be amended during prosecution to require the limitations of the product/apparatus claims. **Failure to do so may result in no rejoinder.** Further, note that the prohibition against double patenting rejections of 35 U.S.C. 121 does not apply where the restriction requirement is withdrawn by the examiner before the patent issues. See MPEP § 804.01.

Conclusion

Any inquiry concerning this communication or earlier communications from the examiner should be directed to TERESA E. STRZELECKA whose telephone number is (571)272-0789. The examiner can normally be reached on M-F (8:30-5:30).

If attempts to reach the examiner by telephone are unsuccessful, the examiner's supervisor, Gary Benzion can be reached on (571) 272-0782. The fax phone number for the organization where this application or proceeding is assigned is 571-273-8300.

Information regarding the status of an application may be obtained from the Patent Application Information Retrieval (PAIR) system. Status information for published applications may be obtained from either Private PAIR or Public PAIR. Status information for unpublished applications is available through Private PAIR only. For more information about the PAIR system, see http://pair-direct.uspto.gov. Should you have questions on access to the Private PAIR system, contact the Electronic Business Center (EBC) at 866-217-9197 (toll-free). If you would like assistance from a USPTO Customer Service Representative or access to the automated information system, call 800-786-9199 (IN USA OR CANADA) or 571-272-1000.

TERESA E STRZELECKA
Primary Examiner
Art Unit 1637

/TERESA E STRZELECKA/
Primary Examiner, Art Unit 1637
January 2, 2015

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 NUMBER	FILING OR 371(C) DATE	FIRST NAMED APPLICANT	ATTY. DOCKET NO./TITLE
14/540,917	11/13/2014	Douglas AMORESE	NUGN-001/01US

CONFIRMATION NO. 7076

POA ACCEPTANCE LETTER

58249
COOLEY LLP
ATTN: Patent Group
1299 Pennsylvania Avenue, NW
Suite 700
Washington, DC 20004

OC000000072703872

Date Mailed: 01/08/2015

NOTICE OF ACCEPTANCE OF POWER OF ATTORNEY

This is in response to the Power of Attorney filed 12/31/2014.

The Power of Attorney in this application is accepted. Correspondence in this application will be mailed to the above address as provided by 37 CFR 1.33.

/hteffera/

Office of Data Management, Application Assistance Unit (571) 272-4000, or (571) 272-4200, or 1-888-786-0101

REMARKS

Restriction

The Office has issued a restriction requirement stating that the application claims two distinct inventions that are related as product and process of use. Specifically, the Office identifies the two inventions as being:

> Group I - claims 1-20, drawn to a method for detecting a duplicate sequencing read from a population of sample sequencing reads; and
>
> Group II - claim 21, drawn to a kit comprising a plurality of adaptors, wherein each adaptor comprises an indexing primer binding site, an indexing site, an identifier site, and a target sequencing primer binding site.

Applicant hereby elects, with traverse, Group I, reading on claims 1-20. The Office alleges there would be a serious search and/or examination burden if restriction were not required because each group alleged requires distinct and non-coextensive search queries. Applicant respectfully disagrees. The claim in Group II is directed to a kit comprising the adaptor recited in Group I. Accordingly, it would not be an undue burden for the Office to search and examine the claim in Group II along with Group I.

In the event the Office maintains the restriction requirement, Applicant reserves the right to pursue claims drawn to the non-elected subject matter of Group II in one or more divisional application(s).

Election of Species

The Office has indicated that the claims are directed to patentably distinct species and has requested an election of species. Applicant must elect a single species from each of the three sets of species:

Species of sequencing the identifier site

A) identifier site is sequenced with the indexing site (claim 3),

B) identifier site is sequenced separately from the indexing site (claim 4),

IDT Exhibit 1018.0115

C) identifier site is sequenced with the target sequence (claim 5),

D) identifier site is sequenced separately from the target sequence (claim 6).

Species of adaptor

E) adaptor comprises from 5' to 3': (i) the indexing primer binding site; (ii) the indexing site; (iii) the identifier site; and (iv) the target sequence primer binding site (claim 7),

F) adaptor comprises from 5' to 3': (i) the indexing primer binding site; (ii) the indexing site; (iii) the target sequence primer binding site; and (iv) the identifier site (claim 8).

Species of nucleic acid fragments

G) genomic DNA fragments (claims 13 and 14),

H) cDNA fragments (claims 13 and 14),

I) RNA fragments (claim 13),

J) DNA/RNA fragments (claim 13).

Applicant hereby elects species A), E) and G). Claims 1-3, 7, and 9-20 encompass the elected species.

Applicant reserves the right to request examination of additional species upon allowance of the elected species in this application, and to pursue claims to the non-elected species in this application in the event that a generic claim is found to be allowable.

Applicant also reserves the right to request rejoinder of claims to non-elected subject matter, upon the allowance of a claim directed to the elected invention. Applicant further reserves the right to file one or more continuation or divisional applications directed to non-elected subject matter.

CONCLUSION

In view of the foregoing, Applicant respectfully submits that submit that this paper fully addresses the Restriction Requirement dated January 7, 2015. However, the Examiner is requested to call the undersigned if any questions or comments arise.

The Director is hereby authorized to charge any appropriate fees under 37 C.F.R. §§1.16, 1.17, and 1.21 that may be required by this paper, and to credit any overpayment, to Deposit Account No. 50-1283.

Dated: March 6, 2015

Respectfully submitted,
COOLEY LLP

COOLEY LLP
ATTN: Patent Group
1299 Pennsylvania Avenue NW, Suite 700
Washington, DC 20004

Tel: (202) 728-7138
Fax: (202) 842-7899

By: /Angie Mah/
Angie S. Mah
Reg. No. 64,902

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

In Re Application of:	AMORESE, Douglas, et al.	Confirmation No.:	7076
Application No.:	14/540,917	Group Art Unit:	1637
Filed:	November 13, 2014	Examiner:	Teresa E. Strzelecka

FOR: **COMPOSITIONS AND METHODS FOR IDENTIFICATION OF A DUPLICATE SEQUENCING READ**

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

RESPONSE TO RESTRICTION REQUIREMENT

This paper is in response to the Restriction Requirement dated January 7, 2015. This response is timely filed before the two-month statutory period for reply ending on March 7, 2015.

Remarks begin on page **2** of this paper.

Reconsideration of the application in view of the following remarks is respectfully requested.

IDT Exhibit 1018.0118

Electronic Acknowledgement Receipt

EFS ID:	21693478
Application Number:	14540917
International Application Number:	
Confirmation Number:	7076
Title of Invention:	COMPOSITIONS AND METHODS FOR IDENTIFICATION OF A DUPLICATE SEQUENCING READ
First Named Inventor/Applicant Name:	Douglas AMORESE
Customer Number:	58249
Filer:	Angie S.Y. Mah
Filer Authorized By:	
Attorney Docket Number:	NUGN-001/01US
Receipt Date:	06-MAR-2015
Filing Date:	13-NOV-2014
Time Stamp:	11:42:55
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	Response to Election / Restriction Filed	NUGN-001_01US_OARtoRestriction_SpeciesRequirement_.pdf	99625 3c4da50339e19780f7024d63e0523843a68a4428	no	4

Warnings:	
Information:	

Total Files Size (in bytes):	99625

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.

Electronic Acknowledgement Receipt

EFS ID:	21917420
Application Number:	14540917
International Application Number:	
Confirmation Number:	7076
Title of Invention:	COMPOSITIONS AND METHODS FOR IDENTIFICATION OF A DUPLICATE SEQUENCING READ
First Named Inventor/Applicant Name:	Douglas AMORESE
Customer Number:	58249
Filer:	Angie S.Y. Mah
Filer Authorized By:	
Attorney Docket Number:	NUGN-001/01US
Receipt Date:	31-MAR-2015
Filing Date:	13-NOV-2014
Time Stamp:	12:03:41
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	Foreign Reference	WO2008093098A2.pdf	2267793 70f0a59bf895afa2f6ec4ac3df5441b07590a177	no	111

Warnings:	
Information:	

2	Foreign Reference	WO2012061832A1.pdf	758620 a73d4bd7edf16a3658dd0a074d9f50b6200f2dba	no	103
Warnings:					
Information:					
3	Foreign Reference	WO2014144092A1.pdf	558648 b075a2ede6e478166f251eef3887421fb38f887f	no	68
Warnings:					
Information:					
4	Non Patent Literature	MYERS_2013.pdf	844636 324d3f99777ec6fe2e74fc11c9833fab1512f61b	no	15
Warnings:					
Information:					
5	Non Patent Literature	PCTUS2014065530__ISR_and_WrittenOpinion.pdf	1037971 b58cb700b1f5390602b5f7353d3e45bdf0d56f40	no	14
Warnings:					
Information:					
6	Transmittal Letter	NUGN-001_01US-IDSTransmittal.pdf	78149 3d269f11858d8841787f783608d134fbab552e9a	no	2
Warnings:					
Information:					
7	Information Disclosure Statement (IDS) Form (SB08)	NUGN-001_01US-SB08.pdf	164578 8d2de668fea53e4b5a38fbf77c1f7d7d1325a1de	no	1
Warnings:					
Information:					
This is not an USPTO supplied IDS fillable form					
		Total Files Size (in bytes):	5710395		

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.

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

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
18 September 2014 (18.09.2014)

WIPO | PCT

(10) International Publication Number
WO 2014/144092 A1

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

(21) International Application Number:
PCT/US2014/028356

(22) International Filing Date:
14 March 2014 (14.03.2014)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
61/801,600 15 March 2013 (15.03.2013) US

(71) Applicant: **NUGEN TECHNOLOGIES, INC.** [US/US]; 201 Industrial Road, Suite 310, San Carlos, CA 94070 (US).

(72) Inventors: **AMORESE, Doug**; 357 South El Monte Avenue, Los Altos, CA 94022 (US). **SCHROEDER, Benjamin, G.**; 1964 Shoreview Avenue, San Mateo, CA 94401 (US). **SCOLNICK, Jonathan**; 451 Kansas Street, #518, San Francisco, CA 94107 (US).

(74) Agents: **HSU, Jennifer** et al.; Wilson Sonsini Goodrich & Rosati, 650 Page Mill Road, Palo Alto, CA 94304-1050 (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, IR, 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, 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, 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))*

[Continued on next page]

(54) **Title**: SEQUENTIAL SEQUENCING

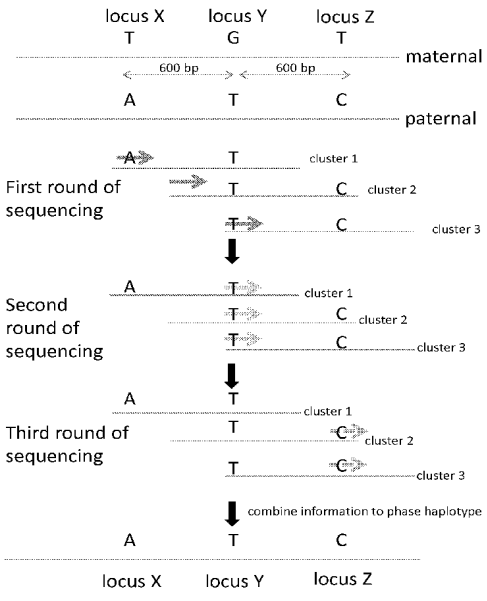


FIG. 3

(57) **Abstract**: Described herein are improved methods, compositions and kits for next generation sequencing (NGS). The methods, compositions and kits described herein enable phasing of two or more nucleic acid sequences in a sample, i.e. determining whether the nucleic acid sequences (which can comprise regions of sequence variation) are located on the same chromosome and/or the same chromosomal fragment. Phasing information can be obtained by performing multiple, successive sequencing reactions from the same immobilized nucleic acid template. The methods, compositions and kits provided herein can be useful, for example, for haplotyping, SNP phasing, or for determining downstream exons in RNA-seq.

WO 2014/144092 A1

IDT Exhibit 1018.0124

— *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))*

SEQUENTIAL SEQUENCING

CROSS-REFERENCE

[0001] This application claims the benefit of U.S. Provisional Application No. 61/801,600, filed March 15, 2013, which application is incorporated herein by reference.

BACKGROUND

[0002] Next generation sequencing (NGS) can be utilized for a wide variety of life science applications, such as, e.g., diagnostics, epidemiology, and forensics. Widely used NGS technologies commonly involve sequencing of short reads (e.g., 100 nt reads) which can then be mapped to a reference genome. While such technologies can be useful for the detection of sequence variants (e.g., mutations), they are also associated with certain limitations. For example, NGS can be used to sequence genomes or genomic regions of diploid organisms. A genome of a diploid organism can comprise sets of chromosome pairs, wherein each chromosome pair comprises a maternal chromosome and a paternal chromosome. For example, a diploid organism can inherent a maternal allele and a paternal allele at a single locus. Thus, alleles detected at a plurality of loci which map to a particular chromosome might exist within a single chromosome (maternal or paternal) of a chromosome pair, or across both chromosomes of a chromosome pair. While many NGS platforms are useful for detecting alleles at a plurality of loci, they currently do not provide phasing and/or haplotype information e.g., to distinguish whether alleles detected at a plurality of loci that map to a particular chromosome are co-located on the same chromosome or are located on separate chromosomes in a chromosome pair. Determining whether a plurality of alleles are co-located on the same (maternal or paternal) chromosome or are located on different chromosomes can be useful for a variety of reasons, as discussed further below.

[0003] The pattern of alleles within each individual chromosome can be referred to as haplotype. Haplotyping can have many diagnostic and clinical applications. For example, two inactivating mutations across different loci within a single gene might be of little or no consequence if present on the same individual chromosome (i.e. chromosome of either maternal or paternal origin), because the other copy of the gene product will remain functional. On the other hand, if one of the inactivating mutations is present in the maternal chromosome and the other in the paternal chromosome, there can be no functional copy of the gene product, which can result in a negative phenotype (non-viability, increased risk for disease and others). Haplotyping can also be used to predict risk or susceptibility to specific

IDT Exhibit 1018.0126

genetic diseases, as many genetic associations are tied to haplotypes. For example, the various haplotypes of the human leukocyte antigen (HLA) system can be associated with genetic diseases ranging from autoimmune disease to cancers.

[0004] Another instance in which phasing information can be useful is distinguishing between functional genes and their non-functional pseudogene counterparts within the genome. One well known functional gene/pseudogene pair is the genes SMN1 and SMN2, which differ in sequence by only five nucleotides over many Kb of sequence, yet one of the nucleotide differences renders the SMN2 gene almost completely non-functional. Using short read sequencing, a mutation may be found in one of the two genes, but unless the mutation happens to occur within the sequencing read that also covers one of the known nucleotide differences between SMN1 and SMN2, it will be difficult to know which of the genes (the functional gene, or the nonfunctional pseudogene) is mutated.

[0005] The present NGS methods can employ short read sequencing to query regions of variable DNA sequence (polymorphisms etc.) interspersed within regions of conserved DNA sequence. As significant blocks of conserved sequence can be interspersed between the variable regions, short read sequencing does not lend itself to phasing analysis. Although methods have been developed to obtain phasing information, these methods (for example, Sanger sequencing and subcloning), can be labor intensive and/or costly.

[0006] There is a need for improved NGS methods that provide phasing information. Such methods can provide a highly parallel platform for performing multiple sequencing reactions from the same immobilized templates. Methods described herein fulfill this need.

SUMMARY OF THE INVENTION

[0007] Provided herein are novel methods, compositions and kits for phasing two or more nucleic acid sequences in a sample. Specifically, in one aspect methods and compositions are provided that can allow for determining whether two or more nucleic acid sequences (typically comprising regions of sequence variation) are located on the same nucleic acid template, chromosome, or chromosomal fragment. The methods and compositions disclosed herein can also be used to distinguish and differentiate between two closely related nucleic acid sequences by compiling and aligning data from sequential sequencing reads.

[0008] The methods, kits, and compositions described herein can employ sequential paired sequencing reads from the same immobilized nucleic acid template. The reads can be generated by successive rounds of priming, sequencing, denaturing and repriming, and the

IDT Exhibit 1018.0127

results from multiple reads originating from the same template can be compiled to obtain phasing information.

[0009] Additionally, the methods, kits, and compositions described herein can employ pools of oligonucleotides which may be used to prime sequencing reactions. Such sequencing reactions can target specific regions of specific nucleic acids for sequencing. These oligonucleotide pools can be used onboard a sequencer to extend the sequencing of nucleic acids that have already undergone a first round of sequencing.

[0010] In one aspect, provided herein is a method for relating multiple nucleic acid sequences (e.g., comprising regions of sequence variation) to the same nucleic acid template. In some embodiments, the method comprises: a) creating a nucleic acid library; b) sequencing the library with an oligonucleotide primer; c) denaturing the extension strand; d) performing a second round of sequencing by introducing a new oligonucleotide primer containing sequence complementary to conserved regions present in some of the nucleic acid templates within the nucleic acid library; e) repeating steps c) and d) as needed; and f) compiling sequencing data from the successive sequencing reads to differentiate between closely related nucleic acid sequences. In some embodiments, the nucleic acid library is a directional nucleic acid library. In some embodiments, the nucleic acid library is not a directional nucleic acid library.

[0011] In some embodiments, the nucleic acid library comprises closely related nucleic acid sequences as inserts. In some embodiments, the conserved regions within the nucleic acid inserts are located adjacent to variable regions. In some embodiments, alignment of multiple variable regions enables differentiating between and/or typing of related transcripts. In some embodiments, alignment of multiple variable regions enables differentiating between and/or typing of related micro-organisms. In some embodiments, the nucleic acid library is a directional nucleic acid library. In some embodiments, the nucleic acid library is not a directional nucleic acid library.

[0012] In another aspect, provided herein is a method for differentiating between closely related nucleic acid sequences (such as genes and pseudogenes) by using specific sets of oligonucleotide primers containing sequence complementary to a common region shared by the closely related sequences. In some embodiments, the method comprises: a) creating a sequencing library with related nucleic acid sequences as inserts; b) sequencing the library with an oligonucleotide primer; c) denaturing the extension strand; d) performing a second round of sequencing by introducing a new oligonucleotide primer containing sequence complementary to conserved regions present in some of the nucleic acid templates within the

IDT Exhibit 1018.0128

nucleic acid library; e) repeating steps c) and d) as needed; and f) compiling sequencing data from the successive sequencing reads to differentiate between closely related nucleic acid sequences.

[0013] Kits for performing any of the methods described herein are provided. Such kits can include reagents, enzymes and platforms for amplification and sequencing of nucleic acids. In one embodiment, a kit is provided comprising: a) an adaptor or several adaptors, b) one or more of oligonucleotide primers, and c) reagents for amplification. In another embodiment, the kit further comprises reagents for sequencing. A kit can include instructions for employing the kit components as well as the use of any other reagent not included in the kit.

[0014] Also provided herein is a method for relating at least two nucleic acid sequences or regions of sequence variation to the same nucleic acid template, the method comprising: (a) creating a nucleic acid library wherein a template is immobilized on a solid support; (b) generating a first sequencing read from the immobilized template with a first oligonucleotide primer, wherein a 3' terminal nucleotide of the first oligonucleotide primer binds to a first binding site in the template; (c) denaturing the first sequencing read from the immobilized template; (d) annealing a second oligonucleotide primer that is complementary to a region or regions within the immobilized template, wherein a 3' terminal nucleotide of the second oligonucleotide primer binds to a second binding site in the template that is more than 5 nt away from the first binding site in the template; (e) generating a second sequencing read with the second oligonucleotide primer; and (f) compiling data from the first sequencing read and the second sequencing read to map reads originating from the template. In some embodiments, the nucleic acid libraries are generated from amplicons. In some embodiments, the amplicons comprise conserved regions which flank a variable region. In some embodiments, alignment of multiple variable regions enables differentiation and/or typing of related transcripts. In some embodiments, alignment of multiple variable regions enables differentiation and/or typing of related micro-organisms. In some embodiments, the libraries are reduced complexity. In some embodiments, the reduced complexity is achieved by target enrichment. In some embodiments, the nucleic acid library comprises closely related nucleic acid sequences as inserts. In some embodiments, the nucleic acid library is a directional nucleic acid library. In some embodiments, the nucleic acid library is not a directional nucleic acid library.

[0015] Also provided herein is a method comprising: (a) coupling at least one nucleic acid template to a sequencing platform; (b) subjecting the at least one nucleic acid template to a first round of sequencing using a first sequencing primer, wherein the first sequencing primer

IDT Exhibit 1018.0129

comprises a 3' terminal nucleotide which anneals to a first binding site in the at least one nucleic acid template, wherein the first round of sequencing produces a first sequencing read of said at least one nucleic acid template; (c) removing said first annealed sequencing primer and said first sequencing read from said at least one nucleic acid template; and (d) subjecting said at least one nucleic acid template to a second round of sequencing using a second sequencing primer, wherein the second sequencing primer comprises a 3' terminal nucleotide which anneals to a second binding site in the at least one nucleic acid template which is more than 5 nt away from the first binding site, wherein the second round of sequencing produces a second sequencing read of the at least one nucleic acid template.

[0016] In some embodiments of any of the foregoing methods, the first and/or the second round of sequencing comprises sequencing by synthesis. In some embodiments, the first and/or the second round of sequencing comprises sequencing by ligation. In some embodiments, the coupling comprises conjugating the at least one nucleic acid template to a solid support of the sequencing platform. In some embodiments, the solid support comprises a bead, a well, an array, or a flow cell. In some embodiments, the at least one nucleic acid template comprises an adaptor oligonucleotide ligated to one or both ends of the at least one nucleic acid template. In some embodiments, the first sequencing primer is selectively hybridizable to a primer binding site of the adaptor oligonucleotide. In some embodiments, the second sequencing primer is not selectively hybridizable to the primer binding site of the adaptor oligonucleotide. In some embodiments, the second sequencing read comprises an informative locus. In some embodiments, the informative locus is a polymorphic locus. In some embodiments, the first sequencing read comprises a first informative locus and the second sequencing read comprises a second informative locus. In some embodiments, the method comprises determining that a first allele detected at the first informative locus and a second allele detected at the second informative locus are in phase. In some embodiments, the second sequencing primer is designed to hybridize to a location that is within x nt upstream of the informative locus, wherein x is a number that is less than an average sequencing read length produced by the first or second round of sequencing. In some embodiments, the removing comprises a denaturing step. In some embodiments, the removing further comprises a washing step. In some embodiments, the method further comprises subjecting the at least one nucleic acid template to one or more additional rounds of sequencing, wherein the one or more additional rounds of sequencing comprise the steps of : (i) removing a previously annealed sequencing primer and previously generated sequencing read from the at least one nucleic acid template; (ii) annealing an additional

IDT Exhibit 1018.0130

sequencing primer to the at least one nucleic acid template; and (iii) subjecting the at least one nucleic acid template to an additional round of sequencing using the additional sequencing primer, thereby producing an additional sequencing read of the at least one nucleic acid template. In some embodiments, the additional sequencing primer comprises a 3' terminal nucleotide which hybridizes to a binding site that is more than 5 nt away from a binding site of the 3' terminal nucleotide of the previously annealed sequencing primer. In some embodiments, a first allele detected by the first sequencing read, a second allele detected by the second sequencing read, and an additional allele detected by the additional sequencing read are determined to be in phase. In some embodiments, the additional sequencing primer and the previously annealed sequencing primer are designed to hybridize to genomic locations that map to a single chromosome. In some embodiments, the genomic locations that map to a single chromosome span a distance that is less than a length of the at least one nucleic acid template. In some embodiments, the genomic locations that map to a single chromosome span a distance that is less than three-quarters of the length of the at least one nucleic acid template. In some embodiments, the genomic locations that map to a single chromosome span a distance that is less than two-thirds of the length of the at least one nucleic acid template. In some embodiments, the genomic locations that map to a single chromosome span a distance that is less than one half of the length of the at least one nucleic acid template. In some embodiments, the genomic locations that map to a single chromosome span a distance that is about a sequencing read length of the sequencing platform. In some embodiments, the at least one nucleic acid template is a member of a sequencing library. In some embodiments, the sequencing library is a whole-genome sequencing library. In some embodiments, the sequencing library is a target-enriched sequencing library. In some embodiments, the sequencing library is a directional sequencing library. In some embodiments, the sequencing library is not a directional sequencing library. In some embodiments, the sequencing library comprises nucleic acid fragments of an average length. In some embodiments, the average length is about 50 bases to about 10 kb. In some embodiments, the first, second, and optionally additional sequencing reads comprise sequence information from a plurality of loci. In some embodiments, phasing of alleles detected at the plurality of loci is determined based upon the first, second, and optionally additional sequencing reads. In some embodiments, the at least one nucleic acid template comprises a plurality of nucleic acid templates. In some embodiments, the at least one nucleic acid template comprises at least 10,000 nucleic acid templates. In some embodiments, the at least one nucleic acid template comprises 3 billion nucleic acid

IDT Exhibit 1018.0131

templates. In some embodiments, the at least one nucleic acid template is coupled to a unique addressable location of the sequencing platform. In some embodiments, the method comprises compiling sequencing reads from the unique addressable location into one data file. In some embodiments, the method comprises determining phasing of a plurality of alleles detected across two or more unique addressable locations. In some embodiments, the two or more unique addressable locations each comprise two or more detected alleles, wherein the two or more detected alleles of each of the two or more unique addressable locations are determined to be in phase if at least one common detected allele is shared by the two or more unique addressable locations. In some embodiments, step (d) of a method disclosed herein comprises subjecting the at least one nucleic acid template to a second round of sequencing using a mixture of second sequencing primers. In some embodiments, the mixture of second sequencing primers comprises a pair of primers, wherein each primer in the primer pair is designed to sequence a forward or reverse complement strand containing a locus. In some embodiments, primers within the mixture are designed and/or selected to have minimal potential to hybridize to the same nucleic acid template.

[0017] In some embodiments of any of the aforementioned methods, the template comprises DNA. In some embodiments, the template comprises RNA. In some embodiments, one or more methods disclosed herein is used to provide phasing information for two or more alleles separated by a distance of over 150 bases, over 300 bases, over 500 bases, over 600 bases, over 700 bases, over 800 bases, over 900 bases, over 1000 bases (1kb), over 5 kb, over 10 kb, over 50 kb, over 100 kb, over 500 kb, over 1000 kb (1 Mb), over 5 Mb, over 10 Mb, over 50 Mb, over 100 Mb, over 150 Mb, over 200 Mb, over 250 Mb, or over 300 Mb.

[0018] Also provided herein is a kit, comprising: (a) one or more oligonucleotide primers for performing a sequencing reaction; and (b) instructions for practicing a method of any of the preceding claims using the one or more oligonucleotide primers.

[0019] Also provided herein is a computer readable medium comprising computer executable code that, upon execution by a computer processor, implements a method comprising: (a) accessing a data file in computer memory that comprises a plurality of sequencing reads corresponding to a first addressable location of a sequencing platform, wherein the sequencing reads are generating by subjecting a first nucleic acid template coupled to the addressable location to a first and at least a second round of sequencing, wherein the first and at least second rounds of sequencing are performed using a first and second sequencing primer, respectively; (b) using the data file, detecting a first detected allele

IDT Exhibit 1018.0132

at a first locus of the nucleic acid template and detecting a second detected allele at a second locus of the nucleic acid template; and (c) determining the first and second detected allele to be in phase based upon said detecting. In some embodiments, the method further comprises: (d) accessing a second data file in computer memory that comprises a plurality of sequencing reads corresponding to a second addressable location of a sequencing platform, wherein the sequencing reads are generated by subjecting a second nucleic acid template coupled to the addressable location to the first and the at least second round of sequencing; (e) using the second data file, detecting a first detected allele at a first locus of the second nucleic acid template and detecting a second detected allele at a second locus of the second nucleic acid template; and (f) determining the first and second detected alleles of the second nucleic acid template to be in phase with the first and second detected alleles of the first nucleic acid template if the first or second detected allele of the second nucleic acid template is shared in common with the first nucleic acid template.

INCORPORATION BY REFERENCE

[0020] All publications, patents, and patent applications mentioned in this specification are herein incorporated by reference to the same extent as if each individual publication, patent, or patent application was specifically and individually indicated to be incorporated by reference.

BRIEF DESCRIPTION OF THE DRAWINGS

[0021] The novel features of the invention are set forth with particularity in the appended claims. A better understanding of the features and advantages of the present invention will be obtained by reference to the following description that sets forth illustrative embodiments, in which the principles of the invention are utilized, and the accompanying drawings of which:

[0022] **FIGS. 1A** and **1B** illustrate the current limitations of NGS sequencing in obtaining phasing information for a plurality of alleles.

[0023] **FIG. 2** depicts an exemplary workflow of a method described herein.

[0024] **FIG. 3** depicts an exemplary embodiment of a method described herein.

[0025] **FIG. 4** depicts another exemplary embodiment of a method described herein.

[0026] **FIG. 5** depicts an exemplary computer system disclosed herein.

[0027] **FIG. 6** depicts sequential sequencing method as applied to 16S microbial rRNA characterization, as described in Example 1.

IDT Exhibit 1018.0133

[0028] **FIG. 7** depicts the use of specific oligonucleotide pools and the generation of mated pairs of sequencing reads to differentiate between two closely related nucleotide sequences, such as a gene/pseudogene pair.

DETAILED DESCRIPTION OF THE INVENTION

[0029] Throughout this application, various embodiments are presented in a range format. It should be understood that the description in range format is merely for convenience and brevity and should not be construed as an inflexible limitation on the scope of the invention. Accordingly, the description of a range should be considered to have specifically disclosed all the possible subranges as well as individual numerical values within that range. For example, description of a range such as from 1 to 6 should be considered to have specifically disclosed subranges such as from 1 to 3, from 1 to 4, from 1 to 5, from 2 to 4, from 2 to 6, from 3 to 6 etc., as well as individual numbers within that range, for example, 1, 2, 3, 4, 5, and 6. This applies regardless of the breadth of the range.

[0030] Unless otherwise specified, terms and symbols of genetics, molecular biology, biochemistry and nucleic acid used herein follow those of standard treatises and texts in the field, e.g. Kornberg and Baker, *DNA Replication,* Second Edition (W.H. Freeman, New York, 1992); Lehninger, *Biochemistry,* Second Edition (Worth Publishers, New York, 1975); Strachan and Read, Human *Molecular Genetics,* Second Edition (Wiley-Liss, New York, 1999); Eckstein, editor, *Oligonucleotides and Analogs: A Practical Approach* (Oxford University Press, New York, 1991); Gait, editor, *Oligonucleotide Synthesis: A Practical Approach* (IRL Press, Oxford, 1984); Sambrook, Fritsch & Maniatis, *Molecular Cloning: A Laboratory Manual*, Fourth Edition (2012); *Oligonucleotide Synthesis* (M. J. Gait, ed., 1984); *Nucleic Acid Hybridization* (B. D. Hames & S. J. Higgins, eds., 1984); *A Practical Guide to Molecular Cloning* (B. Perbal, 1984); and a series, *Methods in Enzymology* (Academic Press, Inc.) and the like.

Definitions

[0031] As used in the specification and claims, the singular forms "a", "an" and "the" can include plural references unless the context clearly dictates otherwise. For example, the term "a primer" can include a plurality of primers, including mixtures thereof.

[0032] The terms "determining", "measuring", "evaluating", "assessing," "assaying," and "analyzing" can be used interchangeably herein to refer to any form of measurement, and include determining if an element is present or not. These terms can include both quantitative and/or qualitative determinations. Assessing may be relative or absolute. "Assessing the

IDT Exhibit 1018.0134

presence of" can include determining the amount of something present, as well as determining whether it is present or absent.

[0033] As used herein, the term "phasing" can refer to the process of determining whether two or more nucleic acid sequences are located on the same nucleic acid template, the same fragment of a chromosome, or the same chromosome.

[0034] As used herein, the term "haplotype" can refer to a haploid genotype, e.g., a combination or set of nucleic acid sequences detected at a plurality of loci which are typically inherited as a unit and are linked, for example during a translocation event. A haplotype can provide a distinctive genetic pattern of an individual. A haplotype can be determined for one locus, several loci, or an entire chromosome depending on the number of recombination events that occur between a given set of loci. The nucleic acid sequences are not limited to any specific type and can include, by way of example only, conserved genetic sequences (e.g., non-variant) or variant genetic sequences (e.g., polymorphisms). For example, single nucleotide polymorphisms (SNPs), short tandem repeats (STRs), including , for example, di-, tri-, and tetra-nucleotide repeats, copy number variations (CNVs), insertions, deletions, inversions, methylations, etc. can be considered variant genetic sequences. Such markers can be found in various publications and databases including, for example, the ATCC short tandem repeat (STR) database, the Human SNP Database, the NCBI dbSNP database, Celera Human SNP database, SNP Database of the Genome Analysis Group, and the like.

[0035] The term "phased alleles" can refer to the distribution of the particular alleles on a nucleic acid template, a single chromosomal region, and/or a single chromosome. Accordingly, the "phase" of two alleles can refer to a characterization or determination of whether the alleles are located on a single chromosome or two separate chromosomes (e.g., a maternally or paternally inherited chromosomes). Unless otherwise stated, "haplotype" and "phased alleles" can be considered synonymous.

[0036] "Nucleotides" and "nt" can be used interchangeably herein to generally refer to biological molecules that can form nucleic acids. Nucleotides can have moieties that contain not only the known purine and pyrimidine bases, but also other bases that have been modified. Such modifications include acylated purines or pyrimidines, methylated purines or pyrimidines, alkylated riboses, or other heterocycles.

[0037] As used herein, the term "oligonucleotide" can refer to a polynucleotide chain, typically less than 200 residues long, most typically between 15 and 100 nucleotides long, but also intended to encompass longer polynucleotide chains. Oligonucleotides can be single-or double-stranded. The terms "oligonucleotide probe" or "probe", as used herein, can refer to

an oligonucleotide capable of hybridizing to a complementary nucleotide sequence. As used herein, the term "oligonucleotide" can be used interchangeably with the terms "primer", "adaptor" and "probe".

[0038] The term "polynucleotide", "nucleic acid", or grammatical equivalents, can refer to two or more nucleotides covalently linked together. A nucleic acid described herein can contain phosphodiester bonds, although in some cases, as outlined below (for example in the construction of primers and probes such as label probes), nucleic acid analogs are included that can have alternate backbones, comprising, for example, phosphoramide (Beaucage et al., Tetrahedron 49(10):1925 (1993) and references therein; Letsinger, J. Org. Chem. 35:3800 (1970); Sprinzl et al., Eur. J. Biochem. 81:579 (1977); Letsinger et al., Nucl. Acids Res. 14:3487 (1986); Sawai et al, Chem. Lett. 805 (1984), Letsinger et al., J. Am. Chem. Soc. 110:4470 (1988); and Pauwels et al., Chemica Scripta 26:141 91986)), phosphorothioate (Mag et al., Nucleic Acids Res. 19:1437 (1991); and U.S. Pat. No. 5,644,048), phosphorodithioate (Briu et al., J. Am. Chem. Soc. 111:2321 (1989), O-methylphosphoroamidite linkages (see Eckstein, Oligonucleotides and Analogues: A Practical Approach, Oxford University Press), and peptide nucleic acid (also referred to herein as "PNA") backbones and linkages (see Egholm, J. Am. Chem. Soc. 114:1895 (1992); Meier et al., Chem. Int. Ed. Engl. 31:1008 (1992); Nielsen, Nature, 365:566 (1993); Carlsson et al., Nature 380:207 (1996), all of which are incorporated by reference. Other analog nucleic acids include those with bicyclic structures including locked nucleic acids (also referred to herein as "LNA"), Koshkin et al., J. Am. Chem. Soc. 120.13252 3 (1998); positive backbones (Denpcy et al., Proc. Natl. Acad. Sci. USA 92:6097 (1995); non-ionic backbones (U.S. Pat. Nos. 5,386,023, 5,637,684, 5,602,240, 5,216,141 and 4,469,863; Kiedrowshi et al., Angew. Chem. Intl. Ed. English 30:423 (1991); Letsinger et al., J. Am. Chem. Soc. 110:4470 (1988); Letsinger et al., Nucleoside & Nucleotide 13:1597 (1994); Chapters 2 and 3, ASC Symposium Series 580, "Carbohydrate Modifications in Antisense Research", Ed. Y. S. Sanghui and P. Dan Cook; Mesmaeker et al., Bioorganic & Medicinal Chem. Lett. 4:395 (1994); Jeffs et al., J. Biomolecular NMR 34:17 (1994); Tetrahedron Lett. 37:743 (1996)) and non-ribose backbones, including those described in U.S. Pat. Nos. 5,235,033 and 5,034,506, and Chapters 6 and 7, ASC Symposium Series 580, "Carbohydrate Modifications in Antisense Research", Ed. Y. S. Sanghui and P. Dan Cook, all of which are incorporated by reference. Nucleic acids containing one or more carbocyclic sugars are also included within the definition of nucleic acids (see Jenkins et al., Chem. Soc. Rev. (1995) pp 169-176, hereby incorporated by reference). Several nucleic acid analogs are described in Rawls, C & E News

IDT Exhibit 1018.0136

Jun. 2, 1997 page 35, hereby incorporated by reference. "Locked nucleic acids" are also included within the definition of nucleic acid analogs. LNAs can be a class of nucleic acid analogues in which the ribose ring is "locked" by a methylene bridge connecting the 2'-O atom with the 4'-C atom. All of these references are hereby expressly incorporated by reference. These modifications of the ribose-phosphate backbone can be done to increase the stability and half-life of such molecules in physiological environments. For example, PNA:DNA and LNA-DNA hybrids can exhibit higher stability and thus can be used in some embodiments. The target nucleic acids can be single stranded or double stranded, as specified, or contain portions of both double stranded or single stranded sequence. Depending on the application, the nucleic acids can be DNA (including, e.g., genomic DNA, mitochondrial DNA, and cDNA), RNA (including, e.g., mRNA and rRNA) or a hybrid, where the nucleic acid contains any combination of deoxyribo- and ribo-nucleotides, and any combination of bases, including uracil, adenine, thymine, cytosine, guanine, inosine, xathanine hypoxathanine, isocytosine, isoguanine, etc.

[0039] As used herein, the terms “hybridization”, “hybridizing” and “annealing” can be used interchangeably and can refer to the pairing of complementary nucleic acids.

[0040] The term “primer”, as used herein, can refer to an oligonucleotide, generally with a free 3’ hydroxyl group, that is capable of hybridizing with a template (such as a target polynucleotide, target DNA, target RNA or a primer extension product) and is also capable of promoting polymerization of a polynucleotide complementary to the template. A primer can contain a non-hybridizing sequence that constitutes a tail of the primer. A primer can still hybridize to a target even though its sequences are not fully complementary to the target.

[0041] “Complementary”, as used herein, can refer to complementarity to all or only to a portion of a sequence. The number of nucleotides in the hybridizable sequence of a specific oligonucleotide primer can be such that stringency conditions used to hybridize the oligonucleotide primer can prevent excessive random non-specific hybridization. The number of nucleotides in the hybridizing portion of the oligonucleotide primer can be at least as great as the defined sequence on the target polynucleotide that the oligonucleotide primer hybridizes to, namely, at least 5, at least 6, at least 7, at least 8, at least 9, at least 10, at least 11, at least 12, at least 13, at least 14, at least 15, at least about 20, or from about 6 to about 10 or 6 to about 12, or 12 to about 200 nucleotides, or about 10 to about 50 nucleotides. In general, the target polynucleotide can be larger than the oligonucleotide primer or primers as described previously.

[0042] The term "genomic sequence", as used herein, can refer to a sequence that occurs in a genome. RNAs can be transcribed from a genome; this term can encompass sequence that exist in the nuclear genome of an organism, as well as sequences that are present in a cDNA copy of an RNA (e.g., an mRNA) transcribed from such a genome.

[0043] The term "target polynucleotide,", "target region", or "target", as used herein, can refer to a polynucleotide of interest under study. A target polynucleotide can contain one or more sequences or loci that are of interest and under study. A target polynucleotide can comprise, for example, a genomic region. The genomic region can comprise all or a portion of a chromosome. The genomic region can comprise a plurality of loci. The genomic region can comprise a plurality of loci for which phasing and/or haplotype information is desired.

[0044] The terms "template", "template strand", "template DNA" and "template nucleic acid" can be used interchangeably herein to refer to a strand of nucleic acid (e.g., DNA) that is copied by an amplification cycle.

[0045] The term "mutation", as used herein, can refer to a change of the nucleotide sequence of a genome. Mutations can involve large regions of a genome (e.g., copy number variation). Mutations can involve small regions of a genome. Examples of mutations involving small sections of a genome include, e.g., point mutations or single nucleotide polymorphisms, multiple nucleotide polymorphisms, insertions (e.g., insertion of one or more nucleotides at a locus), multiple nucleotide changes, deletions (e.g., deletion of one or more nucleotides at a locus), and inversions (e.g., reversal of a sequence of one or more nucleotides).

[0046] The term "locus", as used herein, can refer to a location of a gene, a nucleotide, or a sequence on a chromosome and/or in a genome. A locus can be considered "polymorphic" or to have a "polymorphism" if there exists in a population two or more forms of a nucleotide or sequence at the locus (e.g., if the locus comprises variant genetic sequences). An "allele" of a locus, as used herein, can refer to an alternative form of a nucleotide or sequence at the locus. A "wild-type allele" can refer to an allele that has the highest frequency in a population of subjects. A "mutant allele" can refer to an allele that has a lower frequency that a "wild-type allele". An "informative" locus, as used herein, can refer to a locus which can be used to determine phasing information of two or more alleles. The informative locus can be a polymorphic locus.

[0047] The term "single nucleotide polymorphism", or "SNP", as used herein, can refer to a type of genomic sequence variation resulting from a single nucleotide substitution within a

IDT Exhibit 1018.0138

sequence. "SNP alleles" or "alleles of a SNP" can refer to alternative forms of the SNP at a given locus.

[0048] The term "genotyping", as used herein, can refer to a process of determining differences in the genetic make-up (genotype) of an individual, e.g., by examining the individual's DNA sequence using a biological assay. The biological assay can comprise a sequencing reaction.

[0049] As used herein, a "sample" or "nucleic acid sample" can refer to any substance containing or presumed to contain nucleic acid. The sample can be a biological sample obtained from a subject.

[0050] An "addressable location", as used herein, can refes to a location that can be recorded and/or tracked throughout any of the procedures carried out during one or more methods described herein.

OVERVIEW

[0051] Provided herein are methods, compositions, and kits useful for sequence analysis of nucleic acids. The subject methods, compositions, and kits can be useful for obtaining accurate phasing and/or haplotyping information using NGS technologies. For example, the methods, compositions, and kits provided herein can be used to determine phasing of a plurality of alleles that are separated by a distance that is greater than a distance of an NGS sequence read or paired read. The methods and kits provided herein can be used to obtain phasing information for detected alleles that map to two or more loci. For example, the methods and kits provided herein can be used for determining whether detected alleles that map to two or more loci are located on the same nucleic acid template, on the same chromosome or chromosome region, or on different chromosomes or chromosomal regions in a chromosome pair. The methods and kits can be used to determine a plurality of polymorphic markers at a plurality of loci within a nucleic acid template. The methods and kits described herein can also be used to obtain phasing information for a plurality of alleles that are detected on different nucleic acid templates. One or more phasing methods disclosed herein can be used to provide phasing information for two or more alleles. The two or more alleles may be separated by a distance of over 150 bases, over 300 bases, over 500 bases, over 600 bases, over 700 bases, over 800 bases, over 900 bases, over 1000 bases (1kb), over 5 kb, over 10 kb, over 50 kb, over 100 kb, over 500 kb, over 1000 kb (1 Mb), over 5 Mb, over 10 Mb, over 50 Mb, over 100 Mb, over 150 Mb, over 200 Mb, over 250 Mb, or over 300 Mb.

IDT Exhibit 1018.0139

[0052] The methods and kits described herein can also be used to differentiate between closely related nucleic acid sequences, e.g., to distinguish a gene from a closely related pseudogene. Such methods can be useful, for example, for haplotyping, SNP phasing, determining downstream exons in RNA-seq, and in genetic diagnostics applications. The methods, kits and compositions described herein can utilize sequential sequencing of a nucleic acid template. Altogether, the phasing methods disclosed herein provide an improvement over the existing methods by offering a highly parallel, efficient method for obtaining phasing information.

[0053] **FIGS. 1A** and **1B** illustrate current limitations of NGS sequencing in obtaining phasing information for a plurality of alleles. **FIG. 1A** depicts a chromosome pair comprising a maternal and paternal chromosome obtained from genomic DNA from a subject. The chromosomes comprise a plurality of loci: locus X, locus Y, which is located, e.g., 600 bp downstream of locus X, and locus Z, which is located, e.g., 600 bp downstream of locus Y. The maternal chromosome comprises the following alleles: allele T at locus X (X_T, allele G at locus Y (Y_G), and allele T at locus Z (Z_T). By contrast, the paternal chromosome comprises the following alleles, allele A at locus X (X_A), allele T at locus Y (Y_T), and allele C at locus Z (Z_C). FIG. 1B depicts a current embodiment NGS sequencing method which is unable to determine phasing of alleles at loci X, Y, and Z. Genomic DNA from the subject is fragmented into ~1000 bp DNA fragments. The fragments are spatially separated and clonally amplified into clusters. By way of example only, clusters 1-3 tile overlapping regions which span loci X, Y, and Z. The clusters then undergo next generation sequencing, e.g., on an Illumina platform. Next generation sequencing typically produces sequencing reads from one or both ends of the DNA fragments, due to the sequencing adaptors attached to one or both ends of the fragments, wherein the sequencing adaptors comprise sequencing primer binding sites (not shown in the figure). The sequencing reads can have read lengths of, e.g., 100 nt, so the next generation sequencing in this example does not provide sequencing information across the entire template molecule. In this particular example, paired end sequencing is employed, producing sequencing reads from both ends of the template DNA fragments, each read having a read length of 100 nt (as depicted by thick gray arrows in **FIG. 1B**). Accordingly, the DNA template fragment of cluster 1 comprises X_A and Y_T, however, only X_A is detected by the sequencing reads. The DNA template fragment of cluster 2 comprises Y_T and Z_C, however, only Z_C is detected by the sequencing reads. The DNA template fragment of cluster 3 comprises Y_T and Z_C, however, only Y_T is detected by the sequencing reads. Because sequencing reads from clusters 1-3 only detect a

IDT Exhibit 1018.0140

single allele at a single locus for each cluster, phasing information is not generated for the detected alleles X_A, Y_T, and Z_C, e.g., whether they are located on the same template molecule, the same chromosomal region, or the same chromosome.

[0054] **FIG. 2** depicts an exemplary workflow of a phasing method described herein. The method can comprise a first step 210 of coupling a nucleic acid template to a sequencing platform. The coupling can comprise attaching the nucleic acid template to a solid support of the sequencing platform. Coupling can in some cases include formation of a covalent bond (e.g., formed by ligation), or a noncovalent bond (e.g., formed by hybridization) between the template and the solid support of the sequencing platform. In some cases, coupling does not involve formation of a covalent bond or noncovalent bond between the template and the solid support of the sequencing platform. In some cases, coupling can include associating two entities, e.g., placing an entity, such as a nucleic acid template, in a well. In some cases, the nucleic acid template is coupled to a unique addressable location of the sequencing platform. For example, a plurality of nucleic acid templates can be coupled to a sequencing platform, wherein each of the plurality of templates is coupled to a unique addressable location of the sequencing platform. Coupling of nucleic acid templates to unique addressable locations of a sequencing platform can comprise attaching the templates to the unique addressable locations and/or can comprise confining the templates within the unique addressable locations (e.g., confining to a well). For example, a template can be attached to a solid support (e.g., a bead) and the solid support confined within the unique addressable location, such as, e.g., depositing the bead in a well. The method can further comprise a second step 220 of subjecting the nucleic acid template to a first sequencing reaction using a first sequencing primer. The first sequencing primer can initiate sequencing at a first location on the nucleic acid template. The first sequencing reaction can generate a first sequencing read. The first sequencing read can provide the sequence of a first region of the nucleic acid template. The method can further comprise a second step 230 of subjecting the nucleic acid template to a second sequencing reaction using a second sequencing primer. The second sequencing primer can initiate sequencing at a second location on the nucleic acid template. The second location can be distinct from the first location. In some cases, a 3' terminal nucleotide of the second primer can hybridize to a location that is more than 5 nt away from a binding site of a 3' terminal nucleotide of the first primer. The second sequencing reaction can generate a second sequencing read. The second sequencing read can provide the sequence of a second region of the nucleic acid template which is distinct from the first region of the nucleic acid template. In some embodiments, the nucleic acid template is optionally subjected to one or

IDT Exhibit 1018.0141

more additional rounds of sequencing (depicted as step 240) using additional sequencing primers.

[0055] Nucleic acid templates can be subjected to any number of rounds of sequencing as described herein. For example, nucleic acid templates can be subjected to 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38. 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99, 100, 110, 120, 130, 140, 150, 160, 170, 180, 190, 200, 250, 300, 350, 400, 450, 500, 600, 700, 800, 900, 1000, 2000, 3000, 4000, 5000, 6000, 7000, 8000, 9000, 10000, 100000, or more than 100000 rounds of sequencing according to one or more phasing methods described herein. Nucleic acid templates can be subjected to about 2 to about 8 rounds of sequencing, about 6 to about 20 rounds of sequencing, about 10 to about 100 rounds of sequencing, about 50 to about 500 rounds of sequencing, about 100 to about 1000 rounds of sequencing, about 500 to about 2000 rounds of sequencing, about 1000 to about 10000 rounds of sequencing, or more than about 10000 rounds of sequencing according to one or more phasing methods described herein.

[0056] The rounds of sequencing can produce any number of sequencing reads for the nucleic acid template. For example, about 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99, 100, 110, 120, 130, 140, 150, 160, 170, 180, 190, 200, 250, 300, 350, 400, 450, 500, 600, 700, 800, 900, 1000, 2000, 3000, 4000, 5000, 6000, 7000, 8000, 9000, 10000, 100000, or more than 100000 sequencing reads can be generated for any given nucleic acid template. About 1 to about 8 sequencing reads, about 6 to about 20 sequencing reads, about 10 to about 100 sequencing reads, about 50 to about 500 sequencing reads, about 100 to about 1000 sequencing reads, about 500 to about 2000 sequencing reads, about 1000 to about 10000 sequencing reads, or more than about 10000 sequencing reads can be generated for any given nucleic acid template,

[0057] In some embodiments, the first, second, and/or any additional rounds of sequencing are conducted using a plurality of primers. For example, any round of sequencing designed to sequence a particular locus can comprise use of a pair of primers, wherein each primer in the primer pair is designed to sequence a forward or reverse complement strand containing

IDT Exhibit 1018.0142

the locus. For example, any round of sequencing can comprise sequencing using a mixture of primers. Primers within a mixture can be designed and/or selected to have minimal potential to hybridize to the same nucleic acid template. For example, primers within a mixture can be designed to hybridize to locations that are at least 1 kb, 2 kb, 3kb, 4 kb, 5 kb, 10 kb, 20 kb, 30 kb, 40 kb, 50 kb, 60 kb, 70 kb, 80 kb, 90 kb, 100 kb, 200 kb, 300 kb, 400 kb, 500 kb, 600 kb, 700 kb, 800 kb, 900 kb, 1000 kb (1Mb), 2 Mb, 3 Mb, 4 Mb, 5 Mb, 6 Mb, 7 Mb, 8 Mb, 9 Mb, 10 Mb, , 20 Mb, 30 Mb, 40 Mb, 50 Mb, 60 Mb, 70 Mb, 80 Mb, 90 Mb, 100 Mb, or 200 Mb apart. Primers within a mixture can be designed to hybridize to different chromosomes. A mixture of primers used for any round of sequencing can comprise any number of primers, for example, can comprise 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38. 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99, 100, 110, 120, 130, 140, 150, 160, 170, 180, 190, 200, 250, 300, 350, 400, 450, 500, 600, 700, 800, 900, 1000, 2000, 3000, 4000, 5000, 6000, 7000, 8000, 9000, 10000, 100000, or more than 100000 primers. A mixture of priemrs used for any round of sequencing can comprise about 1 to about 8 primers, about 6 to about 20 primers, about 10 to about 100 primers, about 50 to about 500 primers, about 100 to about 1000 primers, about 500 to about 2000 primers, about 1000 to about 10000 primers, or more than about 10000 primers.

[0058] In some cases, the first, second, and optionally additional sequencing reads can be compiled to produce sequence information for the nucleic acid template. For example, first, second, and optionally additional sequencing reads from a unique addressable location can be compiled to produce sequence information for the nucleic acid template coupled to that unique addressable location.

[0059] **FIG. 3** depicts an exemplary embodiment of a phasing method described herein, and highlights advantages of the subject method for obtaining phasing information for a plurality of alleles. **FIG. 3** depicts the same chromosome pair comprising a maternal and paternal chromosome obtained from genomic DNA from a subject as depicted in **FIG. 1**. As in **FIG. 1**, the chromosomes comprise a plurality of loci: locus X, locus Y, which is located, e.g., 600 bp downstream of locus X, and locus Z, which is located, e.g., 600 bp downstream of locus Y. As in **FIG. 1**, the maternal chromosome comprises the following alleles: allele T at locus X (X_T), allele G at locus Y (Y_G), and allele T at locus Z (Z_T). As in **FIG. 1**, the paternal chromosome comprises the following alleles, allele A at locus X (X_A), allele T at

IDT Exhibit 1018.0143

locus Y (Y_T), and allele C at locus Z (Z_C). As in **FIG. 1**, genomic DNA from the subject is fragmented into ~1000 bp DNA fragments. As in **FIG. 1**, the fragments are spatially separated and clonally amplified into clusters. As in **FIG. 1**, clusters 1-3 tile overlapping regions which span loci X, Y, and Z. The clusters are then subjected to sequential rounds of sequencing according to a phasing method described herein. A first round of sequencing is performed using a first sequencing primer. The first sequencing primer can be designed to selectively hybridize to a sequencing primer binding site of a sequencing adaptor attached to a first end of the cluster templates. The first round of sequencing using the first sequencing primer provides a first set of sequencing reads (dark gray arrows) of a first end of the cluster templates. Because locus X happens to be located near the end of the first end of cluster 1, Allele A_A is detected by the first read of cluster 1. The extension strands resulting from extension of the first sequencing primer are then denatured and removed from the clusters. In a next step, a second round of sequencing is performed using a second sequencing primer. The second sequencing primer can be designed to hybridize to a genomic location upstream of locus Y and to sequence locus Y. Accordingly, the second round of sequencing produces a second set of sequencing reads (medium gray arrows). Because clusters 1, 2, and 3 comprise locus Y, the second set of sequencing reads comprise reads from clusters 1, 2, and 3. The second set of sequencing reads detect allele T at locus Y (Y_T) in clusters 1, 2, and 3. The extension strands resulting from extension of the second sequencing primer are then denatured and removed from the clusters. In a next step, a third round of sequencing is performed using a third sequencing primer. The third sequencing primer can be designed to hybridize to a genomic location upstream of locus Z and to sequence locus Z. Accordingly, the third round of sequencing produces a third set of sequencing reads (stippled medium gray arrows). Because clusters 2 and 3, but not cluster 1, comprise locus Z, the third set of sequencing reads comprise reads from clusters 2 and 3 but not from cluster 1. The third set of sequencing reads detect allele C at locus Z (Z_C) in clusters 2 and 3. Sequence data from the first, second, and third sets of sequencing reads is compiled and used to determine phasing of the detected alleles at loci X, Y, and Z. Because the first and second set of sequencing reads reveal X_A and Y_T in cluster 1, it can be determined that X_A and Y_T are in phase (e.g., on the same nucleic acid fragment). Because the second and third sets of sequencing reads reveal Y_T and Z_C in clusters 2 and 3, it can be determined that Y_T and Z_C are in phase (e.g., on the same nucleic acid fragment). Moreover, because the first, second, and third clusters share a common allele (Y_T), it can be determined that X_A, Y_T, and Z_C are all in

IDT Exhibit 1018.0144

phase. Because X_A, Y_T, and Z_C are all in phase, a haplotype of the subject can be determined to be $X_A Y_T Z_C$.

[0060] **FIG. 4** depicts another exemplary embodiment of a phasing method described herein. **FIG. 4** depicts the same chromosome pair comprising a maternal and paternal chromosome obtained from genomic DNA from a subject as depicted in **FIG. 1**. As in **FIG. 1**, the chromosomes comprise a plurality of loci: locus X, locus Y, which is located, e.g., 600 bp downstream of locus X, and locus Z, which is located, e.g., 600 bp downstream of locus Y. As in **FIG. 1**, the maternal chromosome comprises the following alleles: allele T at locus X (X_T), allele G at locus Y (Y_G), and allele T at locus Z (Z_T). As in **FIG. 1**, the paternal chromosome comprises the following alleles, allele A at locus X (X_A), allele T at locus Y (Y_T), and allele C at locus Z (Z_C). As in **FIG. 1**, genomic DNA from the subject is fragmented into ~1000 bp DNA fragments. As in **FIG. 1**, the fragments are spatially separated and clonally amplified into clusters. As in **FIG. 1**, clusters 1-3 tile overlapping regions which span loci X, Y, and Z. The clusters are then subjected to sequential rounds of sequencing according to a phasing method described herein. A first round of sequencing is performed using a first sequencing primer. The first sequencing primer can be designed to selectively hybridize to a genomic location upstream of locus X and to sequence locus X. The first round of sequencing produces a first set of sequencing reads. Because cluster 1, but not clusters 2 or 3, comprise locus X, the first set of sequencing reads comprises a read from cluster 1 but not from clusters 2 or 3. The first set of sequencing reads detects allele A at locus X(X_A). The extension strands resulting from extension of the first sequencing primer are then denatured and removed from the clusters. In a next step, a second round of sequencing is performed using a second sequencing primer. The second sequencing primer can be designed to hybridize to a genomic location upstream of locus Y and to sequence locus Y. Accordingly, the second round of sequencing produces a second set of sequencing reads (medium gray arrows). Because clusters 1, 2, and 3 all comprise locus Y, the second set of sequencing reads comprise reads from clusters 1, 2, and 3. The second set of sequencing reads detect allele T at locus Y ((Y_T) in clusters 1, 2, and 3. The extension strands resulting from extension of the second sequencing primer are then denatured and removed from the clusters. In a next step, a third round of sequencing is performed using a third sequencing primer. The third sequencing primer can be designed to hybridize to a genomic location upstream of locus Z and to sequence locus Z. Accordingly, the third round of sequencing produces a third set of sequencing reads (stippled medium gray arrows). Because clusters 2 and 3, but not cluster 1, comprise locus Z, the third set of sequencing reads

comprise reads from clusters 2 and 3 but not from cluster 1. The third set of sequencing reads detect allele C at locus Z (Z_C) in clusters 2 and 3. Sequence data from the first, second, and third sets of sequencing reads is compiled and used to determine phasing of the detected alleles at loci X, Y, and Z. Because the first and second set of sequencing reads reveal X_A and Y_T in cluster 1, it can be determined that X_A and Y_T are in phase. Because the second and third sets of sequencing reads reveal Y_T and Z_C in clusters 2 and 3, it can be determined that Y_T and Z_C are in phase. Moreover, because the first, second, and third clusters share a common allele (Y_T), it can be determined that X_A, Y_T, and Z_C are all in phase. Because X_A, Y_T, and Z_C are all in phase, a haplotype of the subject can be determined to be X_A Y_T Z_C.

[0061] Reference will now be made in detail to exemplary embodiments. While the disclosed methods and compositions will be described in conjunction with the exemplary embodiments, it will be understood that these exemplary embodiments are not intended to limit the methods and compositions described herein, but can encompass alternatives, modifications and equivalents, which may be included in the spirit and scope described herein

Exemplary nucleic acid templates

[0062] Nucleic acid templates to be sequenced by a method provided herein can comprise DNA. The DNA can be complex DNA, for example genomic DNA. The DNA can be from a linear or circular genome. The DNA can be plasmid DNA, cosmid DNA, bacterial artificial chromosome (BAC), or yeast artificial chromosome (YAC). The DNA can be from more than one individual or organism. The DNA can be double stranded or single stranded. The input DNA can be part of chromatin. The DNA can be associated with histones. The DNA can comprise one or more chromosomes. For example, if the DNA is from a human, the DNA can comprise one or more of chromosome 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, X, or Y. The DNA can be mitochondrial DNA. The DNA can be cell-free DNA. The DNA can be cDNA. The cDNA can be generated from RNA, e.g., mRNA. Nucleic acid tempaltes to be sequenced by a method provided herein can comprise RNA.

[0063] Nucleic acid templates to be sequenced by a method provided herein can be obtained from and/or can be in a nucleic acid sample. The nucleic acid sample can be obtained from a subject. A subject can be any biological entity of interest from which phasing and/or haplotype information is desired. The subject can be an animal, plant, fungus, insect, bacteria, algae, virus, cell, tissue, and the like. The animal can be human, non-human primate, dog, cat, cow, pig, sheep, guinea pig, hamster, bird, frog, fish, rat, mouse, other

IDT Exhibit 1018.0146

rodent or other animal. A nucleic acid sample can also comprise a mixture of genomes of different species such as host-pathogen, bacterial populations, viral populations, and the like. For example, a DNA sample can comprise DNA (e.g., cDNA) made from a mixture of genomes of different species. Alternatively, the nucleic acid sample can be from a synthetic source.

[0064] The nucleic acid sample can be obtained from a biological sample from the subject. The biological sample can be, e.g., a liquid sample, a solid sample, or a combination thereof. Exemplary liquid samples include, but are not limited to ascites, buccal sample, cavity rinse, cerebrospinal fluid, whole blood, plasma, serum, sweat, tears, sputum, saliva, urine, or organ rinse. The sample can be a cell-free sample. Exemplary cell-free samples include, but are not limited to plasma, serum, plasma, urine, sweat, saliva, sputum, and tears. Exemplary solid biological samples include, e.g., tissue biopsy, feces, and the like. A biological sample can also comprise cultured cells or constituents of culture cells (including, e.g., conditioned medium resulting from the growth of cells in cell culture medium, recombinant cells and cell components). The sample can be a single cell. Nucleic acid samples can be obtained from biological samples by any means known to those of skill in the art. For example, nucleic acid can be extracted from a biological sample using liquid extraction (e.g, Trizol, DNAzol) techniques. Nucleic acid can also be extracted using commercially available kits. Nucleic acids can be concentrated by, e.g., centrifugation and/or precipitation. Nucleic acids can be purified by binding to a selective membrane and then by elution. In some cases, gentle handling steps may be taken to avoid shearing of chromosomal or intact DNA during the extraction process. Exemplary gentle handling strategies can include, e.g., pipetting with a wide-bore pipette tip and avoidance of vigorous pipetting and/or vortexing. In some cases, gentle handling steps are unnecessary for preparation of nucleic acid samples.

[0065] The nucleic acid sample can be fragmented by any means known in the art or otherwise described herein. The nucleic acids can be fragmented to result in a population of nucleic acid fragments of a desired length. The nucleic acid fragments can be double stranded or single stranded fragments. The length of the nucleic acid fragments can be any length. A nucleic acid fragment to be sequenced by a method described herein can have a length that is less than 300 Mb, less than 200 Mb, less than 100 Mb, less than 50 Mb, less than 40 Mb, less than 30 Mb, less than 20 Mb, less than 10 Mb, less than 5 Mb, less than 3 Mb, less than 2 Mb, less than 1 Mb (1000 kb), less than 900 kb, less than 800 kb, less than 700 kb, less than 600 kb, less than 500 kb, less than 400 kb, less than 300 kb, less than 200 kb, less than 100 kb (100,000 bases), less than 100,000 bases, less than 90,000 bases, less

than 80,000 bases, less than 70,000 bases, less than 60,000 bases, less than 50,000 bases, less than 40000 bases, less than 30,000 bases, less than 20,000 bases, less than 10,000 bases, less than 9000 bases, less than 8000 bases, less than 7000 bases, less than 6000 bases, less than 5000 bases, less than 4000 bases, less than 3000 bases, less than 2000 bases, less than 1500 bases, less than 1400 bases, less than 1300 bases, less than 1200 bases, less than 1100 bases, less than 1000 bases, less than 900 bases, less than 800 bases, less than 700 bases, less than 600 bases, less than 500 bases, less than 400 bases, less than 300 bases, less than 250 bases, less than 200 bases, less than 150 bases, less than 100 bases, less than 50 bases, less than 40 bases, less than 30 bases, less than 20 bases, or less than 10 bases. In some cases, a nucleic acid fragment to be sequenced by a method described herein has a length that is about 10 bases, about 20 bases, about 30 bases, about 40 bases, about 50 bases, about 100 bases, about 150 bases, about 200 bases, about 250 bases, about 300 bases, about 400 bases, about 500 bases, about 600 bases, about 700 bases, about 800 bases, about 900 bases, about 1000 bases, about 1100 bases, about 1200 bases, about 1300 bases, about 1400 bases, about 1500 bases, about 1600 bases, about 1700 bases, about 1800 bases, about 1900 bases, about 2000 bases, about 3000 bases, about 4000 bases, about 5000 bases, about 6000 bases, about 7000 bases, about 8000 bases, about 9000 bases, about 10000 bases, about 20000 bases, about 30000 bases, about 40000 bases, about 50000 bases, about 60000 bases, about 70000 bases, about 80000 bases, about 90000 bases, about 100000 bases (100 kb), about 200 kb, about 300 kb, about 400 kb, about 500 kb, about 600 kb, about 700 kb, about 800 kb, about 900 kb, about 1000 kb (1Mb), about 2 Mb, about 3 Mb, about 5 Mb, about 10 Mb, about 10 Mb, about 30 Mb, about 40 Mb, about 50 Mb, about 100 Mb, about 200 Mb, or about 300 Mb. In some cases, a nucleic acid fragment to be sequenced by a method described herein has a length that is more than about 10 bases, more than about 20 bases, moer than about 30 bases, more than about 40 bases, more than about 50 bases, more than about 100 bases, more than about 150 bases, more than about 200 bases, more than about 250 bases, more than about 300 bases, more than about 400 bases, more than about 500 bases, more than about 600 bases, more than about 700 bases, more than about 800 bases, more than about 900 bases, more than about 1000 bases, more than about 1100 bases, more than about 1200 bases, more than about 1300 bases, more than about 1400 bases, more than about 1500 bases, more than about 1600 bases, more than about 1700 bases, more than about 1800 bases, more than about 1900 bases, more than about 2000 bases, more than about 3000 bases, more than about 4000 bases, more than about 5000 bases, more than about 6000 bases, more than about 7000 bases, more than about 8000 bases, more than about 9000 bases, more than about 10000 bases, more than about

IDT Exhibit 1018.0148

20000 bases, more than about 30000 bases, more than about 40000 bases, more than about 50000 bases, more than about 60000 bases, more than about 70000 bases, more than about 80000 bases, more than about 90000 bases, more than about 100000 bases (100 kb), more than about 200 kb, more than about 300 kb, more than about 400 kb, more than about 500 kb, more than about 600 kb, more than about 700 kb, more than about 800 kb, more than about 900 kb, more than about 1000 kb (1Mb), more than about 2 Mb, more than about 3 Mb, more than about 5 Mb, more than about 10 Mb, more than about 10 Mb, more than about 30 Mb, more than about 40 Mb, more than about 50 Mb, more than about 100 Mb, more than about 200 Mb, or more than about 300 Mb. A nucleic acid fragment to be sequenced by a method described herein can have a length that is between about 10 to about 100 bases, about 50 to about 200 bases, about 100 to about 250 bases, about 150 to about 500 bases, about 300 to about 700 bases, about 400 to about 800 bases, about 600 to about 900 bases, about 800 to about 1000 bases, about 200 to about 1000 bases, about 500 bases to about 10000 bases, or about 2000 bases to about 100000 bases. In some embodiments, the nucleic acid fragment has a length that is over 9000 bases. For example, the nucleic acid fragments can be about 0.9 to about 2 kb, about 1.5 to about 4 kb, about 2 to about 6 kb, about 5 to about 8 kb, or about 7 to about 10 kb, about 10 to about 100 kb, about 50 kb to about 1 Mb, about 1 Mb to about 50 Mb, about 10 Mb to about 300 Mb. One of skill in the art will readily understand that the length of the nucleic acid templates to be sequenced by a method disclosed herein can be any length. Length of nucleic acid templates to be sequenced by a method disclosed herein can vary depending on the sequencing platform used in practicing a method disclosed herein. In some embodiments, the nucleic acid sample is not fragmented.

[0066] The nucleic acid sample can be enriched for target polynucleotides of interest. Target polynucleotides can comprise polynucleotides that map to genomic regions of interest. A genomic region of interest can be any genomic region from which phasing and/or haplotype information is desired. The genomic region of interest can comprise a region of a chromosome. The genomic region of interest can comprise a whole chromosome. The chromosome can be a diploid chromosome. In a human genome, for example, the diploid chromosome can be any of chromosomes 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23. In some cases, the chromosome can be an X or Y chromosome. In some cases, the genomic region of interest comprises a portion of a chromosome. A genomic region of interest can be of any length. The genomic region of interest can have a length that is between, e.g., about 1 to about 10 bases, about 5 to about 50 bases, about 10 to about 100 bases, about 70 to about 300 bases, about 200 bases to about 1000 bases (1 kb), about 700

IDT Exhibit 1018.0149

bases to about 2000 bases, about 1 kb to about 10 kb, about 5 kb to about 50 kb, about 20 kb to about 100 kb, about 50 kb to about 500 kb, about 100 kb to about 2000 kb (2 Mb), about 1 Mb to about 50 Mb, about 10 Mb to about 100 Mb, about 50 Mb to about 300 Mb. For example, a genomic region of interest can be over 1 base, over 10 bases, over 20 bases, over 50 bases, over 100 bases, over 200 bases, over 400 bases, over 600 bases, over 800 bases, over 1000 bases (1kb), over 1.5 kb, over 2 kb, over 3 kb, over 4 kb, over 5 kb, over 10 kb, over 20 kb, over 30 kb, over 40 kb, over 50 kb, over 60 kb, over 70 kb, over 80 kb, over 90 kb, over 100 kb, over 200 kb, over 300 kb, over 400 kb, over 500 kb, over 600 kb, over 700 kb, over 800 kb, over 900 kb, over 1000 kb (1Mb), over 2 Mb, over 3 Mb, over 4 Mb, over 5 Mb, over 6 Mb, over 7 Mb, over 8 Mb, over 9 Mb, over 10 Mb, , over 20 Mb, over 30 Mb, over 40 Mb, over 50 Mb, over 60 Mb, over 70 Mb, over 80 Mb, over 90 Mb, over 100 Mb, or over 200 Mb.

[0067] A genomic region of interest can comprise one or more informative loci. An informative locus can be a polymorphic locus, e.g., comprising two or more alleles. In some cases, the two or more alleles comprise a minor allele. In particular embodiments, genomic regions of interest comprise one or more HLA loci.

[0068] Enrichment of target polynucleotides can be by any means known in the art or otherwise described herein. For example, nucleic acids can be enriched using target-specific primers, or by hybridization to target-selective probes. Target-selective probes can comprise a nucleic acid sequence that is selectively hybridizable to a target polynucleotide of interest. The target-selective probes can comprise a capture moiety which enables capture by a capture reagent. Exemplary capture moiety/reagents include, e.g., biotin/avidin, streptavidin, neutravidin, digoxigenin/wheat germ agglutinin, histidine/cobalt, iron. Nucleic acids can be enriched by a target-selective sequencing library preparation method. Target-selective sequencing library preparation kits and technology are commercially available from, e.g., NuGEN, Agilent, Illumina, or Nimblegen. Target-selective sequencing library preparation methods from Agilient are described in, e.g., US Patent No. 7,867,703 and US Patent Application Publication No. 20120289426, which are hereby incorporated by reference. Target-selective sequencing library preparation methods from Illumina are described in, e.g., US Patent Application Pub No. 20070141604, hereby incorporated by reference. Target-selective sequencing library preparation methods from Nimblegen are described in, e.g., US Patent Application Pub. No. 20090105081, hereby incorporated by reference. Target-selective sequencing library preparation methods from NuGEN are described in, e.g., US Patent Application Pub. No. 20130231253, hereby incorporated by reference. Other

IDT Exhibit 1018.0150

exemplary target-selective sequencing library preparation methods are described in US Patent Application Publication No. 20120003657, hereby incorporated by reference. Enrichment of target genomic regions can, in some cases, comprise isolation of intact chromosomes or chromosomal regions from a biological sample. In some embodiments, the nucleic acid sample is not enriched for target polynucleotides, e.g., represents a whole genome.

[0069] The nucleic acid template to be sequenced can be a member of a sequencing library. Sequencing libraries can comprise any number of library members. Sequencing libraries can comprise 2, 10, 50, 100, 500, 1000, 5000, 10000, 50000, 100000, 500000, 1000000, 5000000, 10000000, 50000000, 100000000, 500000000, 1000000000, 2000000000, 3000000000, 4000000000, 5000000000, 6000000000, 7000000000, 8000000000, 9000000000, 10000000000, or more than 10000000000 members. In some embodiments, nucleic acid libraries are sequenced according to one or more methods described herein. The sequencing library can be prepared from a nucleic acid sample by any means known to those of skill in the art or otherwise described herein. In some embodiments, library preparation comprises attachment of one or more adaptor oligonucleotides to one or more ends of nucleic acid fragments. In some cases, adaptor oligonucleotides are attached to both ends of nucleic acid fragments. The term "adaptor", as used herein, can refer to an oligonucleotide. In some cases, the attachment of an adaptor to a nucleic acid template can enable the generation of amplification-ready products of the nucleic acid template. An adaptor can have various designs. In some embodiments, an adaptor comprises a primer binding site for a platform-specific sequencing primer. In some embodiments, an adaptor comprises a barcode sequence. The barcode sequence can be a unique sequence of nucleotides that can encode information about an assay. A barcode sequence can, for example, encode information relating to the identity of an interrogated allele, identity of a target region of interest, identity of a sample, identity of a subject, or any combination thereof.

[0070] Attachment of adaptor oligonucleotides to nucleic acid fragments can be by any means known in the art, including, e.g., hybridization, amplification, ligation, and the like. Various ligation processes and reagents are known in the art and can be useful for attachment of adaptor oligonucleotides to nucleic acid fragments. For example, blunt ligation can be employed. Similarly, a single dA nucleotide can be added to the 3'-end of the double-stranded DNA product, by a polymerase lacking 3'-exonuclease activity and can anneal to an adaptor comprising a dT overhang (or the reverse). This design allows the hybridized components to be subsequently ligated (e.g., by T4 DNA ligase). Other ligation strategies and the corresponding reagents and known in the art and kits and reagents for carrying out

IDT Exhibit 1018.0151

efficient ligation reactions are commercially available (e.g, from New England Biolabs, Roche).

[0071] In some cases, it is useful to preserve information about the direction of single-stranded nucleic acid molecules while generating double-stranded polynucleotides. This may be accomplished by creation of a strand specific, e.g., directional sequencing library. The term "strand specific" or "directional", as used herein, can refer to the ability to differentiate in a double-stranded polynucleotide between the original template strand and the strand that is complementary to the original template strand. In some cases, one of the strands of the double-stranded polynucleotide is synthesized so that it has at least one modified nucleotide incorporated into it along the entire length of the strand. In some embodiments, the incorporation of the modified nucleotide marks the strand for degradation or removal. In some embodiments, the methods provided herein contemplate construction of directional nucleic acid libraries as described in pending US Application, Serial No. 13/643,056, titled COMPOSITIONS AND METHODS FOR DIRECTIONAL NUCLEIC ACID AMPLIFICATION AND SEQUENCING, Serial No. 13/643,056, hereby incorporated by reference in its entirety. In some cases, information about the direction of single-stranded nucleic acid molecules is not preserved while generating double-stranded polynucleotides. In some cases, the sequencing library is not a directional sequencing library.

[0072] Nucleic acid library members can optionally be amplified prior to sequencing. Methods of amplification are well known in the art. In some embodiments, the amplification is exponential, e.g. in the enzymatic amplification of specific double stranded sequences of DNA by a polymerase chain reaction (PCR). In other embodiments the amplification method is linear. In other embodiments the amplification method is isothermal. In some embodiments, the amplification comprises digital PCR. The digital PCR can be droplet digital PCR. In some embodiments, the amplification comprises single primer isothermal amplification (SPIA). In some embodiments, the amplification comprises reverse transcription PCR. In some embodiments, the amplification comprises real-time PCR. In some embodiments, the amplification comprises quantitative PCR (Q-PCR).

[0073] Nucleic acid library members can be clonally amplified prior to sequencing. Clonal amplification can involve generation of identical copies of individual nucleic acid templates in a nucleic acid library. Nucleic acid templates can be clonally amplified by any means known to those of skill in the art or otherwise described herein. In some cases, clonal amplification can comprise spatial separating individual nucleic acid template molecules and performing amplification of the separated nucleic acid template molecules. In certain

IDT Exhibit 1018.0152

instances, clonal amplification can comprise emulsion PCR (454 Life Sciences, Life Technologies). Emulsion PCR can comprise attachment of an individual nucleic acid template to a bead or particle. Beads and/or particles can be isolated in a water-in-oil emulsion, wherein droplets of the water-in-oil emulsion comprise a bead or particle with a template molecule attached thereon, and reaction components for carrying out an amplification reaction. Amplification of templates within the emulsion can result in clonal amplification of the templates. Clonal amplification can comprise bridge amplification (Illumina, Inc.). Bridge amplification is described, e.g, in US Patent No. 7,985,565, which is hereby incorporated by reference. Clonal amplification can comprise template walking (Life Technologies), described in US Patent Application Publication No. 20120156728, hereby incorporated by reference.

Primers

[0074] Primers can be oligonucleotides that are employed in an extension reaction by a polymerase along a polynucleotide template, such as in PCR or cDNA synthesis, for example. An oligonucleotide primer can be a synthetic polynucleotide that is single stranded, containing a sequence at its 3'-end that is capable of hybridizing with a sequence of the target polynucleotide. In some cases, the 3' region of the primer that hybridizes with the target nucleic acid has at least 80%, 90%, 95%, or 100% complementarity to a sequence or primer binding site.

[0075] Primers can be sufficiently long to prime synthesis of an extension strand along a template. Primers used in one or more methods described herein can have a length. The length can be any length, e.g., about 5 to about 15 nt, about 10 to about 20 nt, about 15 to about 30 nt, about 20 to about 40 nt, about 30 to about 60 nt, about 50 to about 100 nt, or more than about 100 nt.

[0076] Primers used in one or more methods described herein can include primers designed to hybridize to a portion or all of a sequencing adaptor. In some cases, primers described herein are designed to hybridize to a primer-binding sequence of a sequencing adaptor. Exemplary adaptors are described herein. In some cases, a first sequencing primer to be used in a sequential sequencing method (e.g., primer or primers to be used in a first round of sequencing) is designed to hybridize to a portion or all of a sequencing adaptor.

[0077] Primers used in one or more methods can include primers designed to sequence one or more loci from which phasing information is desired. The one or more loci can comprise one or more informative loci. For example, in some cases, the first sequencing primer, second sequencing primer (e.g., primer or primers to be used in a first or second round of

IDT Exhibit 1018.0153

sequencing, respectively), or any subsequent sequencing primer used in a sequential sequencing method is designed to sequence an informative locus.

[0078] A primer designed to sequence an informative locus can be designed to hybridize to a genomic location that is a distance upstream of the informative locus to be sequenced. The distance can be of any length. A skilled artisan will readily appreciate that the distance can vary depending on the read length of the sequencing technology utilized in the sequential sequencing method. In some cases, the distance is shorter than a read length of the sequencing technology utilized in the sequential sequencing method. The primer designed to sequence an informative locus can be designed to hybridize to a genomic location that is about, e.g., 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99, 100, 110, 120, 130, 140, 150, 160, 170, 180, 190, 200, 250, 300, 350, 400, 450, 500, 600, 700, 800, 900, 1000, 2000, 3000, 4000, 5000, 6000, 7000, 8000, 9000, 10000, 100000, or more than 100000 nt upstream of the informative locus. The primer designed to sequence an informative locus can be designed to hybridize to a genomic location that is about 1 to about 10, about 5 to about 50, about 20 to about 100, about 40 to about 200, about 80 to about 500, about 100 to about 1000, about 500 to about 10000, about 5000 to about 100000, or more than 100000 nt upstream of the informative locus.

[0079] In some cases, the primer designed to sequence an informative locus can be designed to hybridize to a conserved region of the genome. The conserved region can be a region that does not comprise a polymorphic locus, e.g., is substantially not polymorphic. In some embodiments, the conserved region exhibits a polymorphism (e.g., sequence variant) at a population frequency that is less than 5%, less than 4%, less than 3%, less than 2%, less than 1%, less than 0.9%, less than 0.8%, less than 0.7%, less than 0.6%, less that 0.5%, less than 0.4%, less than 0.3%, less than 0.2%, less than 0.1%, less than 0.05%, less than 0.01%, less than 0.05%, less than 0.001%, less than 0.005%, less than 0.0001%, %, less than 0.0005%, less than 0.00001%,%, less than 0.00005%, less than 0.000001%, less than 0.000005%, less than 0.0000001%, less than 0.0000005%, or less than 0.00000001%. In some cases, the primer designed to sequence an informative locus can be designed to hybridize to an unconserved region of the genome.

[0080] Primers used in a first, second, or additional round of sequencing can hybridize to distinct locations on the nucleic acid template. For example, a first primer used in a first

IDT Exhibit 1018.0154

round of sequencing may comprise a 3' terminal nucleotide that hybridizes to a first location of a nucleic acid template. A second primer used in a second round of sequencing may comprise a 3' terminal nucleotide that hybridizes to a second location that is more than 5 nt away from the first location of the nucleic acid template (e.g., hybridizes to a location that is at least 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38. 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99, 100, 110, 120, 130, 140, 150, 160, 170, 180, 190, 200, 250, 300, 350, 400, 450, 500, 600, 700, 800, 900, 1000, 2000, 3000, 4000, 5000, 6000, 7000, 8000, 9000, 10000, 100000, or more than 100000 nt away from the first location of the nucleic acid template. In some cases, a third primer used in a third round of sequencing may comprise a 3' terminal nucleotide that hybridizes to a third location that is more than 5 nt away from the second location of the nucleic acid template (e.g., hybridizes to a nucleotide that is at least 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38. 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99, 100, 110, 120, 130, 140, 150, 160, 170, 180, 190, 200, 250, 300, 350, 400, 450, 500, 600, 700, 800, 900, 1000, 2000, 3000, 4000, 5000, 6000, 7000, 8000, 9000, 10000, 100000, or more than 100000 nt away from the second location of the nucleic acid template. Any primer used in a subsequent round of sequencing can have a 3' terminal nt that hybridizes to a location of the nucleic acid template that is at least 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38. 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99, 100, 110, 120, 130, 140, 150, 160, 170, 180, 190, 200, 250, 300, 350, 400, 450, 500, 600, 700, 800, 900, 1000, 2000, 3000, 4000, 5000, 6000, 7000, 8000, 9000, 10000, 100000, or more than 100000 nt away from a binding site of the 3' terminal nt of the previous sequencing primer used in the previous round of sequencing.

[0081] Primers used in a first, second, or additional round of sequencing can initiate sequencing at distinct locations on the nucleic acid template. For example, a first primer used in a first round of sequencing may initiate sequencing at a first location of a nucleic acid template. A second primer used in a second round of sequencing may initiate sequencing at a

IDT Exhibit 1018.0155

second location that is more than 5 nt away from the first location of the nucleic acid template (e.g., initiates sequencing at a location that is at least 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99, 100, 110, 120, 130, 140, 150, 160, 170, 180, 190, 200, 250, 300, 350, 400, 450, 500, 600, 700, 800, 900, 1000, 2000, 3000, 4000, 5000, 6000, 7000, 8000, 9000, 10000, 100000, or more than 100000 nt away from the first location of the nucleic acid template. In some cases, a third primer used in a third round of sequencing may initiate sequencing at a third location that is more than 5 nt away from the second location of the nucleic acid template (e.g., initiates sequencing at a location that is at least 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99, 100, 110, 120, 130, 140, 150, 160, 170, 180, 190, 200, 250, 300, 350, 400, 450, 500, 600, 700, 800, 900, 1000, 2000, 3000, 4000, 5000, 6000, 7000, 8000, 9000, 10000, 100000, or more than 100000 nt away from the second location of the nucleic acid template. Any primer used in a subsequent round of sequencing can initiate sequencing at a location of the nucleic acid template that is more than 5 nt away from the location of initiated sequencing for the previous round of sequencing, e.g., at least 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99, 100, 110, 120, 130, 140, 150, 160, 170, 180, 190, 200, 250, 300, 350, 400, 450, 500, 600, 700, 800, 900, 1000, 2000, 3000, 4000, 5000, 6000, 7000, 8000, 9000, 10000, 100000, or more than 100000 nt away from the location of initiated sequencing for the previous round of sequencing.

[0082] In some embodiments, standard or universal sequencing primers are used. For example, primers designed to hybridize to a primer binding site of a sequencing adaptor can be used. A primer can be a polyA or polyT oligonucleotide. In some embodiments, sequence-specific primers that hybridize to a conserved region or conserved regions within the nucleic acid inserts in the sequencing library are used. In some embodiments, the sequence-specific primers are designed to hybridize to conserved regions adjacent to regions of variable sequence within the nucleic acid inserts, thereby enabling differentiating between

IDT Exhibit 1018.0156

closely related sequences. In some embodiments, a set of oligonucleotide primers that hybridize to sequences shared in closely related sequences, such as gene/pseudogene pairs, are used.

[0083] Primers can be prepared by any suitable method. Methods for preparing oligonucleotides of specific sequence are known in the art, and include, for example, cloning and restriction of appropriate sequences or direct chemical synthesis. Chemical synthesis methods may include, for example, the phosphodiester method disclosed by Brown et al., 1979, Methods in Enzymology 68:109, the phosphotriester method described by Narang et al., 1979, Methods in Enzymology 68:90, the diethylphosphoramidate method disclosed in Beaucage et al., 1981, Tetrahedron Letters 22:1859, and the solid support method disclosed in U.S. Pat. No. 4,458,066, all of which publications are hereby incorporated by reference. Primers and/or reporter probes can also be obtained from commercial sources such as Amersham Pharmacia Biotech, Operon Technologies, Sigma, IDT Technologies, and Life Technologies. Computer programs can also be used to design primers, including but not limited to Array Designer Software (Arrayit Inc.), NetPrimer, Oligonucleotide Probe Sequence Design Software for Genetic Analysis (Olympus Optical Co.), Primer3, and DNAsis from Hitachi Software Engineering.

Methods of Sequencing

[0084] The methods provided herein contemplate sequential sequencing of nucleic acid templates. Sequencing methods are also well known in the art. Sequencing can be performed by a sequencing platform. The sequencing platform can be a next-generation sequencing (NGS) platform. In some cases, the NGS platform can sequence clonally amplified DNA templates or single DNA molecules in a massively parallel fashion (e.g. as described in Volkerding et al. Clin Chem 55:641-658 [2009]; Metzker M Nature Rev 11:31-46 [2010]), hereby incorporated by reference. Sequential sequencing methods provided herein can also employ Sanger sequencing.

[0085] The next-generation sequencing platform can be a commercially available platform. Commercially available platforms useful for performing a method described herein include, e.g., platforms for sequencing-by-synthesis, ion semiconductor sequencing (ChemFET sequencing), reversible dye terminator sequencing, pyrosequencing, sequencing by ligation, single-molecule sequencing, and sequencing by hybridization.

[0086] Platforms for sequencing by synthesis are available from, e.g., 454 Life Sciences, Illumina, Helicos Biosciences, and Qiagen. Illumina platforms can include, e.g., Illumina's Solexa platform, Illumina's Genome Analyzer, Illumina's MiSeq, Illumina's HiSeq,

IDT Exhibit 1018.0157

Illumina's NextSeq, and are described in Gudmundsson et al (Nat. Genet. 2009 41:1122-6), Out et al (Hum. Mutat. 2009 30:1703-12) and Turner (Nat. Methods 2009 6:315-6), U.S. Patent Application Publication Nos. 20080160580 and 20080286795, U.S. Patent Nos. 6,306,597, 7,115,400, and 7,232,656, which are hereby incorporated by reference. 454 Life Science platforms include, e.g., the GS Flex and GS Junior, and are described in U.S. Patent No. 7,323,305, hereby incorporated by reference. Platforms for ion semiconductor sequencing include, e.g., the Ion Torrent Personal Genome Machine (PGM) and are described in U.S. Patent No. 7,948,015, hereby incorporated by reference. Platforms for pryosequencing include the GS Flex 454 system and are described in U.S. Patent Nos. 7,211,390; 7,244,559; 7,264,929, hereby incorporated by reference. Platforms for sequencing by ligation include, e.g., the SOLiD sequencing platform and are described in U.S. Patent No. 5,750,341, hereby incorporated by reference. Platforms for single-molecule sequencing include the SMRT system from Pacific Bioscience and the Helicos True Single Molecule Sequencing platform.

[0087] For example, a sequencing technique that can be used in one or more phasing methods described herein is the method commercialized by Illumina, as described, e.g., in U.S. Pat. Nos. 5,750,341; 6,306,597; and 5,969,119. Such method can employ cluster amplification of library members onto a flow cell and a sequencing-by-synthesis approach. In a round of sequencing, cluster-amplified library members can be subjected to repeated cycles of polymerase-directed single base extension. Single-base extension can involve incorporation of reversible-terminator dNTPs, each dNTP labeled with a different removable fluorophore. The reversible-terminator dNTPs can be 3' modified to prevent further extension by the polymerase. After incorporation of the dNTP, the incorporated nucleotide can be identified by fluorescence imaging. Following fluorescence imaging, the fluorophore can be removed and the 3' modification can be removed resulting in a 3' hydroxyl group, thereby allowing another cycle of single base extension.

[0088] In one embodiment of a phasing method employing Illumina sequencing methodology, a sequencing library is prepared from a nucleic acid sample obtained from a subject. In some cases, the sequencing library is a directional sequencing library. In some cases, the sequencing library is not a directional sequencing library. Single-stranded nucleic acids from the library can optionally be amplified, for example, by PCR, prior to coupling to the Illumina sequencing platform. To couple template nucleic acids to the Illumina platform, the nucleic acids can be denatured into single-stranded templates. The single-stranded templates can be randomly attached to the inside surface of flow-cell channels. In some

IDT Exhibit 1018.0158

embodiments the single-stranded templates are randomly attached to unique addressable locations of the flow-cell. Unlabeled nucleotides can be added to initiate solid-phase bridge amplification to produce dense clusters of clonally amplified templates. In some embodiments, each cluster is coupled to a unique addressable location of the flow-cell. To initiate a first round of sequencing, labeled reversible terminator dNTPs, a first sequencing primer, and DNA polymerase can be added to the Illumina flow cell. The clusters can be subjected to repeated cycles of polymerase-directed single base extension and label detection as described above, thereby producing a first sequencing read from one or more clusters. Upon completion of the first round of sequencing, the extended strands and sequencing primers can be denatured from the cluster templates and removed, e.g., by washing. To initiate a second round of sequencing, labeled reversible terminator dNTPS, a second sequencing primer, and DNA polymerase can be added to the Illumina flow cell. The clusters can then be subjected to repeated cycles of polymerase-directed single base extension and label detection as described above, thereby producing a second sequencing read from one or more clusters. Subsequent rounds of sequencing can optionally be performed by removing the extended strands and primers from the preceding round of sequencing, using subsequent primers that are designed to hybridize, for example, downstream or upstream of the preceding sequencing primer. Sequence reads for the first, second, and optionally additional rounds of sequencing can be compiled for each cluster. The compiling can be used to generate sequence information for the clonally amplified templates. Sequence information for the clonally amplified templates can be used to determine phasing of one or more alleles detected by the sequencing.

[0089] In some embodiments, one or more phasing methods described herein employ sequencing by ligation methods commercialized by Applied Biosystems (e.g., SOLiD sequencing). Library preparation for SOLiD sequencing can comprise ligation of adaptors attached to the 5' and 3' ends of the fragments to generate a fragment library. Alternatively, internal adaptors can be introduced by ligating adaptors to the 5' and 3' ends of the fragments, circularizing the fragments, digesting the circularized fragment to generate an internal adaptor, and attaching adaptors to the 5' and 3 ' ends of the resulting fragments to generate a mate-paired library. Next, nucleic acid templates can be clonally amplified on beads via PCR. Following PCR, the templates can be denatured. Beads can be enriched for beads with extended templates. The clonally amplified templates on the beads can be subjected to a 3' modification that permits bonding to a glass slide. The beads can be deposited on a glass slide for sequencing by ligation. In some embodiments, each bead is deposited onto a unique

IDT Exhibit 1018.0159

addressable location of the glass slide. A round of SOLiD sequencing can comprise the following steps: (1) hybridization of a first sequencing primer to the clonally amplified templates. The first sequencing primer can be designed to hybridize to a sequencing primer binding site on one or more adaptors attached to the template nucleic acids. (2) Following hybridization, a set of four fluorescently labeled di-base probes can compete for ligation to the sequencing primer. Specificity of the di-base probe can be achieved by interrogating every 1st and 2nd base in each ligation reaction. (3) Cycles of ligation, detection and cleavage can be performed with the number of cycles determining the eventual read length. (4) Following a series of ligation cycles, the extension product can be removed and the template can be reset with a primer complementary to the n-1 position for a second round of ligation cycles. For SOLiD sequencing, one round of sequencing can comprise five rounds of primer reset and ligation cycles as described above, wherein each subsequent primer hybridizes to an n-1 position of the adaptor sequencing.

[0090] In one embodiment of a phasing method described herein employing SOLiD sequencing methodology, a sequencing library is prepared from a nucleic acid sample obtained from a subject using a library preparation method for SOLiD sequencing as described herein. In some cases, the sequencing library is a directional sequencing library. In some cases, the sequencing library is not a directional sequencing library. The sequencing library members can be deposited onto unique addressable locations of the SOLiD platform as described herein. A first round of SOLiD sequencing can be performed as described herein. Following the first round of sequencing, the extension product can be removed and a second set of sequencing primers can be used for a second round of SOLiD sequencing. The second round of SOLiD sequencing can comprise multiple cycles of primer hybridization and ligation as described above. The second set of sequencing primers can comprise a primer "primer A" that is designed to hybridize to a genomic location residing 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, or more than 50 nucleotides upstream of an informative locus "A" as well as 2, 3, 4, 5, or more additional sequencing primers that are designed to hybridize to an n-1 genomic location, n-2 genomic location, n-3 genomic location, n-4 genomic location, n-5 genomic location, etc., of "Primer A" Following a second round of SOLiD sequencing, the extension product can be removed and a third set of sequencing primers can be used for a third round of SOLiD sequencing. The third set of sequencing primers can comprise a primer "primer B" that is designed to hybridize to a genomic location residing 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21,

IDT Exhibit 1018.0160

22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, or more than 50 nucleotides upstream of an informative locus "B"., as well as 2, 3, 4, 5, or more additional sequencing primers that are designed to hybridize to an n-1 location, n-2 location, n-3 location, n-4 location, n-5 location, etc., of "Primer B". The informative locus "B" can be located downstream or upstream of informative locus "A". The third round of SOLiD sequencing can comprise multiple cycles of primer hybridization and ligation as described above. Additional rounds of SOLiD sequencing can be performed to sequence further downstream informative loci, as described above. Sequence reads for the first, second, and optionally additional rounds of sequencing can be compiled for each addressable location. The compiling can be used to generate sequence information for the clonally amplified templates. Sequence information for the clonally amplified templates can be used to determine phasing of one or more alleles detected by the sequencing.

[0091] In other embodiments, one or more phasing methods disclosed herein employ sequencing by synthesis using the methods commercialized by 454/Roche Life Sciences, including but not limited to the methods and apparatus described, e.g., in Margulies et al., *Nature(2005)* 437:376-380 (2005); and U.S. Pat. Nos. 7,244,559; 7,335,762; 7,211,390; 7,244,567; 7,264,929; and 7,323,305, hereby incorporated by reference. 454 sequencing library preparation can comprise ligation of adaptors to the ends of nucleic acid fragments. The adaptors can serve as primers for amplification and can provide a sequencing primer binding site for sequencing-by-synthesis. One of the adaptors can comprise a capture reagent, e.g., a biotin. The nucleic acid fragments can be attached to capture beads, e.g., streptavidin-coated beads. The fragments attached to the beads can be PCR amplified within droplets of an oil-water emulsion, resulting in multiple copies of clonally amplified nucleic acid fragments on each bead. Following 454 library preparation, library members can be coupled to a 454 sequencing platform by capture of beads in wells of the 454 platform, which can be pico-liter sized. In some cases, a well captures one bead. Each well can accordingly be an addressable location of the 454 platform. Pyrosequencing can be performed on each DNA fragment in parallel. Pyrosequencing can comprise hybridization of a sequencing primer to a sequencing primer binding site on an adaptor, followed by extension of the primer in a pyrosequencing reaction. During the pyrosequencing reaction, dTNPs are added such that only one out of four of the possible A/T/C/G nucleotides are added and available at a time. Pyrosequencing can detect release of pyrophosphate (PPi) upon nucleotide incorporation. PPi can be converted to ATP by ATP sulfurylase in the presence of adenosine 5' phosphosulfate. Luciferase can use ATP to convert luciferin to oxyluciferin, thereby

IDT Exhibit 1018.0161

generating a light signal that is detected. A detected light signal can be used to identify the incorporated nucleotide. The intensity of the light signal can be used to determine how many bases are incorporated. Following detection of the light signal, unincorporated dNTPs can be removed and the next one out of four possible A/T/C/G nucleotides can be added.

[0092] In one embodiment of a phasing method employing pyrosequencing methodology (e.g., 454 sequencing), a sequencing library is prepared from a nucleic acid sample obtained from a subject. In some cases, the sequencing library is a directional sequencing library. In some cases, the sequencing library is not a directional sequencing library. Sequencing library members can be coupled to addressable locations of a pyrosequencing platform (e.g., a 454 sequencing platform) as described above. A first round of pyrosequencing can be performed as described above, generating a first sequencing read from the library members. Following the first round of pyrosequencing, the extension strands and primers are removed from the library members, e.g., by denaturation and washing. A second round of sequencing can then be performed by hybridization of a second sequencing primer to one or more library members. The second sequencing primer can be designed to hybridize to a location upstream of an informative locus "A" and to sequence informative locus "A". A second pyrosequencing reaction can then be performed as described above. Subsequent removal of the second sequencing primer and additional rounds of pyrosequencing can be performed to sequence additional informative loci. The additional informative loci can be located downstream or upstream of locus "A". Sequence reads for the first, second, and optionally additional rounds of sequencing can be compiled for each addressable location. The compiling can be used to generate sequence information for the clonally amplified templates. Sequence information for the clonally amplified templates can be used to determine phasing of two or more alleles detected by the sequencing.

[0093] In other embodiments, one or more phasing methods disclosed herein may employ the sequencing methods commercialized by Helicos BioSciences Corporation (Cambridge, Mass.) as described, e.g., in U.S. Patent Application Ser. No. 11/167,046, and U.S. Pat. Nos. 7,501,245; 7,491,498; 7,276,720; and in U.S. Patent Application Publication Nos. US20090061439; US20080087826; US20060286566; US20060024711; US20060024678; US20080213770; and US20080103058, hereby incorporated by reference. For example, a method can employ the true Single-Molecule Sequencing technology by Helicos Biosciences Corporation. tSMS library preparation can comprise ligation of a polyA adaptor to 3 ' end of DNA fragments. The adapted fragments can then be coupled to a Helicos tSMS sequencer by the following steps: (1) fragments are hybridized to poly-T oligonucleotides immobilized on

IDT Exhibit 1018.0162

the tSMS flow cell; (2) the flow cell is loaded into an instrument, e.g., HeliScope™ sequencer. Each fragment can be hybridized to a unique addressable location on the tSMS flow cell. A laser can illuminate the surface of the flow cell, revealing the position of each template. A CCD camera can map the position of the templates on the flow cell surface. tSMS sequencing can then proceed by subjecting library members to repeated cycles of polymerase-directed single base extension. A round of tSMS sequencing reaction can commence by introducing a DNA polymerase and a fluorescently labeled nucleotide. The polymerase can incorporate the labeled nucleotides to the primer in a template directed manner. The polymerase and unincorporated nucleotides can be removed. The templates that have directed incorporation of the fluorescently labeled nucleotide can be discerned by imaging the flow cell surface. After imaging, a cleavage step can remove the fluorescent label, and the process can be repeated with other fluorescently labeled nucleotides until a desired read length is achieved. Sequence information can be collected with each nucleotide addition step.

[0094] In one embodiment of a phasing method employing the Helicos tSMS methodology, a tSMS sequencing library is prepared as described above from a nucleic acid sample obtained from a subject. In some cases, the sequencing library is a directional sequencing library. In some cases, the sequencing library is not a directional sequencing library. The sequencing library members can be coupled to unique addressable locations on the tSMS platform as described herein, e.g., by hyridization to polyT oligonucleotides immobilized on the flow cell. Reverse complement complement strands of the hybridized library members can be synthesized, thus creating immobilized templates for sequencing. The original library members hybridized to the flowcell can be denatured and/or removed. A first round of tSMS sequencing can be performed on immobilized templates as described above, generating a first sequencing read from the immobilized templates. Following the first round of tSMS sequencing, the first sequencing reads can be denatured and/or removed. The immobilized templates can be retained on the flow cell. A second round of tSMS sequencing can then be performed by hybridization of a sequencing primer to the immobilized templates. The second sequencing primer can be designed to hybridize to a location upstream of an informative locus "A" and to sequence informative locus "A". A second tSMS sequencing reaction can then be performed as described above. Subsequent denaturation and/or removal of the sequencing primer and sequence reads, followed by additional rounds of tSMS sequencing can be performed on the immobilized templates to sequence additional informative loci. The additional informative loci can be located

IDT Exhibit 1018.0163

downstream or upstream of locus "A". Sequence reads for the first, second, and optionally additional rounds of sequencing can be compiled for each addressable location. The compiling can be used to generate sequence information for the templates. Sequence information for the templates can be used to determine phasing of one or more alleles detected by the sequencing.

[0095] Another example of a sequencing technique that can be used in the methods described herein is semiconductor sequencing provided by Ion Torrent (e.g., using the Ion Personal Genome Machine (PGM)). Ion Torrent technology can use a semiconductor chip with multiple layers, e.g., a layer with micro-machined wells, an ion-sensitive layer, and an ion sensor layer. Sequencing library preparation for the Ion Torrent platform can involves ligation of two distinct adaptors at both ends of a DNA fragment. Individual library members can be attached to beads, as described herein. The library members can then be clonally amplified by, e.g., emulsion PCR. The clonally amplified library members can be coupled to an Ion Torrent sequencing platform, for example, by introducing individual beads comprising the clonally amplified library members into wells of the Ion Torrent platform. Each well can be considered a unique addressable location of the Ion Torrent platform. A round of sequencing can comprise annealing a first sequencing primer to the library members. The first sequencing primer can be designed to hybridize to a sequencing primer binding region of an Ion Torrent library adaptor. To initiate sequencing of the nucleic acids on the beads, one type of deoxyribonucleotide (e.g., dATP, dCTP, dGTP, or dTTP) can be introduced into the wells. When one or more nucleotides are incorporated by DNA polymerase, protons (hydrogen ions) can be released in the well, which can be detected by the ion sensor. The semiconductor chip can then be washed and the process can be repeated with a different deoxyribonucleotide. A plurality of nucleic acids can be sequenced in the wells of a semiconductor chip. The semiconductor chip can comprise chemical-sensitive field effect transistor (chemFET) arrays to sequence DNA (for example, as described in U.S. Patent Application Publication No. 20090026082). Incorporation of one or more triphosphates into a new nucleic acid strand at the 3' end of the sequencing primer can be detected by a change in current by a chemFET. An array can have multiple chemFET sensors.

[0096] In one embodiment of a phasing method described herein employing Ion Torrent methodology, an Ion Torrent sequencing library is prepared as described above from a nucleic acid sample obtained from a subject. In some cases, the sequencing library is a directional sequencing library. In some cases, the sequencing library is not a directional sequencing library. Sequencing library members can be coupled to unique addressable

IDT Exhibit 1018.0164

locations of the Ion Torrent platform as described herein. A first round of Ion Torrent sequencing can be performed on library members as described above, generating a first sequencing read from the library members. Upon completion of the first round of sequencing, the extended strands and sequencing primers can be denatured from the templates and removed, e.g., by washing. A second round of sequencing can be performed using a second sequencing primer. The second sequencing primer can be designed to hybridize to a location upstream of an informative locus "A" and to sequence informative locus "A". A second sequencing reaction can then be performed as described above. Subsequent removal of the second sequencing primer and additional rounds of sequencing can be performed to sequence additional informative loci. The additional informative loci can be located downstream or upstream of locus "A". Sequence reads for the first, second, and optionally additional rounds of sequencing can be compiled for each addressable location. The compiling can be used to generate sequence information for the clonally amplified templates. Sequence information for the clonally amplified templates can be used to determine phasing of one or more alleles detected by the sequencing.

[0097] Sequential sequencing by any of the means described herein can produce one or more sequencing reads for nucleic acid templates. The sequencing reads can have a length which may vary depending on the particular sequencing technology utilized. NGS platforms can provide sequence reads that vary in size from 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, tens to hundreds, or thousands of base pairs. In some embodiments of the method described herein, the sequence reads are about 1 base long, about 2 bases long, about 3 bases long, about 4 bases long, about 5 bases long, about 6 bases long, about 7 bases long, about 8 bases long, about 9 bases long, about 10 bases long, about 12 bases long, about 14 bases long, about 16 bases long, about 18 bases long, about 20 bases long, about 25 bases long, about 30 bases long, about 35 bases long, about 40 bases long, about 45 bases long, about 50 bases long, about 55 bases long, about 60 bases long, about 65 bases long, about 70 bases long, about 75 bases long, about 80 bases long, about 85 bases long, about 90 bases long, about 95 bases long, about 100 bases long, about 110 bases long, about 120 bases long, about 130, about 140 bases long, about 150 bases long, about 200 bases long, about 250 bases long, about 300 bases long, about 350 bases long, about 400 bases long, about 450 bases long, about 500 bases long, about 600 bases long, about 700 bases long, about 800 bases long, about 900 bases long, about 1000 bases, about 2000 bases, about 3000 bases, about 4000 bases, about 5000 bases, about 6000 bases, about 7000 bases, about 8000 bases, about 9000 bases, about 10,000 bases long, about 20,000 bases long, about 30,000 bases long, about 40,000 bases long, about

50,000 bases long, about 60,000 bases long, about 70,000 bases long, about 80,000 bases long, about 90,000 bases long, about 100,000 bases, about 150,000 bases long, about 200,000 bases long, about 250,000 bases long, or more than 250,000 bases long.

[0098] In some cases, a sequencing read is at least 1 base long, at least 2 bases long, at least 3 bases long, at least 4 bases long, at least 5 bases long, at least 6 bases long, at least 7 bases long, at least 8 bases long, at least 9 bases long, at least 10 bases long, at least 12 bases long, at least 14 bases long, at least 16 bases long, at least 18 bases long, at least 20 bases long, at least 25 bases long, at least 30 bases long, at least 35 bases long, at least 40 bases long, at least 45 bases long, at least 50 bases long, at least 55 bases long, at least 60 bases long, at least 65 bases long, at least 70 bases long, at least 75 bases long, at least 80 bases long, at least 85 bases long, at least 90 bases long, at least 95 bases long, at least 100 bases long, at least 110 bases long, at least 120 bases long, at least 130 bases long, at least 140 bases long, at least 150 bases long, at least 200 bases long, at least 250 bases long, at least 300 bases long, at least 350 bases long, at least 400 bases long, at least 450 bases long, at least 500 bases long, at least 600 bases long, at least 700 bases long, at least 800 bases long, at least 900 bases long, at least 1000 bases long, at least 10,000 bases long, at least 20,000 bases long, at least 30,000 bases long, at least 40,000 bases long, at least 50,000 bases long, at least 60,000 bases long, at least 70,000 bases long, at least 80,000 bases long, at least 90,000 bases long, at least 100,000 bases, at least 150,000 bases long, at least 200,000 bases long, or more than 200,000 bases long.

[0099] A sequencing read can be between about 1 to about 10 bases long, between about 5 to about 50 bases long, between about 20 to about 100 bases long, between about 50 to about 200 bases long, between about 100 to about 500 bases long, between about 200 to about 1000 bases long, between about 500 to about 2000 bases long, between about 1000 to about 5000 bases long, between about 2000 to about 10000 bases long, between about 5000 to about 50000 bases long, abetween about 10000 to about 100000 bases long, or more than 100000 bases long.

Sequence analysis

[00100] Sequencing reads (e.g., sequence tags) can be aligned to assemble a partial or complete map of the subject's genome. Sequence tags can be aligned (e.g., mapped), for example, to a reference genome. The reference genome can be a human reference genome. The human reference genome can be the NCBI36/hg18 sequence, available on the world wide web at genome.ucsc.edu/cgi-bin/hgGateway?org=Human&db=hgl 8&hgsid=166260105). Sequencing reads can be mapped using an alternative source of

IDT Exhibit 1018.0166

public sequence information. Other sources of public sequence information include GenBank, dbEST, dbSTS, EMBL (the European Molecular Biology Laboratory), and the DDBJ (the DNA Databank of Japan). In some cases, the reference genome can be an artificial target sequence genome comprising polymorphic target sequences.

[00101] Mapping of the sequence tags can be achieved, for example, by aligning the sequence of the tag to a sequence of the reference genome. Such methods can be used to map a sequencing read to a genomic region, a chromosomal region, or a chromosome. A number of computer algorithms can be used for aligning sequences. Such algorithms include, without limitation, BLAST (Altschul et al., 1990), BLITZ (MPsrch) (Sturrock & Collins, 1993), BOWTIE (Langmead et al, Genome Biology 10:R25.1-R25.10 [2009]), FASTA (Person & Lipman, 1988), SAMtools (SAMtools, Bioinformatics, 2009, 25(16):2078-9), and ELAND (Illumina, Inc., San Diego, CA, USA). In one embodiment, sequence reads produced by Illumina sequencing are processed by bioinformatic alignment analysis for the Illumina Genome Analyzer, which uses the Efficient Large-Scale Alignment of Nucleotide Databases (ELAND) software. Additional software includes the Burroughs-Wheeler block sorting compression procedure which involves block sorting or preprocessing to make compression more efficient.

[00102] Reads from the sequential rounds of sequencing can be compiled for each sequenced template to assemble complete or partial sequence information of the template. Reads from the sequential rounds of sequencing can be compiled according to unique addressable locations of the sequencing platform. For example, all reads generated from a unique addressable location of a sequencing platform can be compiled into a data file. Reads can be compiled with aid of a computer processor. Compiled reads from a unique addressable location can be used to determine partial or complete sequence information for a nucleic acid template (either single template or clonally amplified template). The partial or complete sequence information can comprise information over a longer distance than an average read length of the sequencing platform. Because the sequence information is over a longer distance than an average read length of the sequencing platform, the probability that the sequence information comprises information over two or more informative loci can be increased.

[00103] Compiled reads from a unique addressable location may comprise sequence information for two or more detected alleles. Such compiled reads can be used to determine that the two or more detected alleles are in phase, since the detected alleles are located on the sample template or clonally amplified template. Alleles detected across two or more

IDT Exhibit 1018.0167

addressable locations can also be determined to be in phase, if the two or more addressable locations each share at least one detected allele in common. Alleles detected across a set of addressable locations comprising any number of addressable locations can be determined to be in phase, if the addressable locations in the set share at least one detected allele in common with at least one other addressable location in the set.

[00104] Compiling of reads, determination of phasing information, and/or haplotype analysis can be performed with the aid of a computer system. In some embodiments, the computer system executes instructions contained in a computer-readable medium. The computer system can comprise a processor. The processor can be associated with one or more controllers, calculation units, and/or other units of a computer system, or implanted in firmware. In some embodiments, one or more steps of the method are implemented in hardware. In some embodiments, one or more steps of the method are implemented in software. Software routines may be stored in any computer readable memory unit such as flash memory, RAM, ROM, magnetic disk, laser disk, or other storage medium as described herein or known in the art. Software may be communicated to a computing device by any known communication method including, for example, over a communication channel such as a telephone line, the internet, a wireless connection, or by a transportable medium, such as a computer readable disk, flash drive, etc. The one or more steps of the methods described herein may be implemented as various operations, tools, blocks, modules and techniques which, in turn, may be implemented in firmware, hardware, software, or any combination of firmware, hardware, and software. When implemented in hardware, some or all of the blocks, operations, techniques, etc. may be implemented in, for example, an application specific integrated circuit (ASIC), custom integrated circuit (IC), field programmable logic array (FPGA), or programmable logic array (PLA).

[00105] **FIG. 5** depicts an exemplary computer system adapted to implement one or more methods provided herein. The system **500** includes a central computer server **501** that is programmed to implement exemplary methods described herein. The server **501** includes a central processing unit (CPU, also "processor") **505** which can be a single core processor, a multi core processor, or plurality of processors for parallel processing. The server **501** also includes memory **510** (e.g. random access memory, read-only memory, flash memory); electronic storage unit **515** (e.g. hard disk); communications interface **520** (e.g. network adaptor) for communicating with one or more other systems; and peripheral devices **525** which may include cache, other memory, data storage, and/or electronic display adaptors. The memory **510**, storage unit **515**, interface **520**, and peripheral devices **525** are in

communication with the processor **505** through a communications bus (solid lines), such as a motherboard. The storage unit **515** can be a data storage unit for storing data. The server **501** is operatively coupled to a computer network ("network") **530** with the aid of the communications interface **520**. The network **530** can be the Internet, an intranet and/or an extranet, an intranet and/or extranet that is in communication with the Internet, a telecommunication or data network. The network **530** in some cases, with the aid of the server **501**, can implement a peer-to-peer network, which may enable devices coupled to the server **501** to behave as a client or a server.

[00106] The storage unit **515** can store files, such as subject reports, and/or communications with the caregiver, sequencing data, data about individuals, or any aspect of data generated by one or more phasing methods disclosed herein.

[00107] The server can communicate with one or more remote computer systems through the network **530**. The one or more remote computer systems may be, for example, personal computers, laptops, tablets, telephones, Smart phones, or personal digital assistants.

[00108] In some situations the system **500** includes a single server **501**. In other situations, the system includes multiple servers in communication with one another through an intranet, extranet and/or the Internet.

[00109] The server **501** can be adapted to store sequencing information, and/or information on a subject, such as, for example, polymorphisms, mutations, patient history of the subject, demographic data and/or other information of potential relevance. Such information can be stored on the storage unit **515** or the server **501** and such data can be transmitted through a network.

[00110] Methods as described herein can be implemented by way of machine (or computer processor) executable code (or software) stored on an electronic storage location of the server **501**, such as, for example, on the memory **510**, or electronic storage unit **515**. During use, the code can be executed by the processor **505**. In some cases, the code can be retrieved from the storage unit **515** and stored on the memory **510** for ready access by the processor **505**. In some situations, the electronic storage unit **515** can be precluded, and machine-executable instructions are stored on memory **510**. Alternatively, the code can be executed on a second computer system **540**.

[00111] Aspects of the systems and methods provided herein, such as the server **501**, can be embodied in programming. Various aspects of the technology may be thought of as "products" or "articles of manufacture" typically in the form of machine (or processor) executable code and/or associated data that is carried on or embodied in a type of machine

readable medium. Machine-executable code can be stored on an electronic storage unit, such memory (e.g., read-only memory, random-access memory, flash memory) or a hard disk. "Storage" type media can include any or all of the tangible memory of the computers, processors or the like, or associated modules thereof, such as various semiconductor memories, tape drives, disk drives and the like, which may provide non-transitory storage at any time for the software programming. All or portions of the software may at times be communicated through the Internet or various other telecommunication networks. Such communications, for example, may enable loading of the software from one computer or processor into another, for example, from a management server or host computer into the computer platform of an application server. Thus, another type of media that may bear the software elements includes optical, electrical, and electromagnetic waves, such as used across physical interfaces between local devices, through wired and optical landline networks and over various air-links. The physical elements that carry such waves, such as wired or wireless likes, optical links, or the like, also may be considered as media bearing the software. As used herein, unless restricted to non-transitory, tangible "storage" media, terms such as computer or machine "readable medium" can refer to any medium that participates in providing instructions to a processor for execution.

[00112] Hence, a machine readable medium, such as computer-executable code, may take many forms, including but not limited to, tangible storage medium, a carrier wave medium, or physical transmission medium. Non-volatile storage media can include, for example, optical or magnetic disks, such as any of the storage devices in any computer(s) or the like, such may be used to implement the system. Tangible transmission media can include: coaxial cables, copper wires, and fiber optics (including the wires that comprise a bus within a computer system). Carrier-wave transmission media may take the form of electric or electromagnetic signals, or acoustic or light waves such as those generated during radio frequency (RF) and infrared (IR) data communications. Common forms of computer-readable media therefore include, for example: a floppy disk, a flexible disk, hard disk, magnetic tape, any other magnetic medium, a CD-ROM, DVD, DVD-ROM, any other optical medium, punch cards, paper tame, any other physical storage medium with patterns of holes, a RAM, a ROM, a PROM and EPROM, a FLASH-EPROM, any other memory chip or cartridge, a carrier wave transporting data or instructions, cables, or links transporting such carrier wave, or any other medium from which a computer may read programming code and/or data. Many of these forms of computer readable media may be involved in carrying one or more sequences of one or more instructions to a processor for execution.

[00113] The computer readable medium can comprise computer executable code. The computer executable code can implement a method comprising: (a) accessing a data file in computer memory that comprises a plurality of sequencing reads corresponding to a first addressable location of a sequencing platform, wherein the sequencing reads are generating by subjecting a first nucleic acid template coupled to the addressable location to a first and at least a second round of sequencing, wherein the first and at least second rounds of sequencing are performed using a first and second sequencing primer, respectively; (b) using said data file, detecting a first detected allele at a first locus of the nucleic acid template and detecting a second detected allele at a second locus of the nucleic acid template; and (c) determining the first and second detected allele to be in phase based upon said detecting. The method executed by the computer executable code can further comprise the steps of (d) accessing a second data file in computer memory that comprises a plurality of sequencing reads corresponding to a second addressable location of a sequencing platform, wherein the sequencing reads are generating by subjecting a second nucleic acid template coupled to the addressable location to the first and the at least second round of sequencing; (e) using the second data file, detecting a first detected allele at a first locus of the second nucleic acid template and detecting a second detected allele at a second locus of the second nucleic acid template; and (f) determining the first and second detected alleles of the second nucleic acid template to be in phase with the first and second detected alleles of the first nucleic acid template if the first or second detected allele of the second nucleic acid template is shared in common with the first nucleic acid template.

Kits

[00114] Any of the compositions described herein may be comprised in a kit. In a non-limiting example, the kit, in a suitable container, comprises: an adaptor or several adaptors, one or more of oligonucleotide primers and reagents for ligation, primer extension and amplification. The kit may also comprise means for purification, such as a bead suspension, and nucleic acid modifying enzymes. Kits can further comprise instructions for practicing any of the phasing methods described herein.

Products

[00115] Products based on the methods provided herein may be commercialized by the Applicants under the Ovation® family. Ovation is a registered trademark of NuGEN Technologies, Inc.

[00116] The information gained from the phasing of alleles in a genome as provided by practicing the methods described herein can be of great utility in research and discovery as

IDT Exhibit 1018.0171

well as for clinical, diagnostic, forensic, archeological, geneological, and/or epidemiological applications. For example, haplotype information can be used to aid in the diagnosis, prognosis, and/or theranosis of a disease. For other example, haplotype information of a potential tissue or organ donor and organ recipient can aid in the assessment of HLA compatibilities of the donor and recipient for decreasing transplant rejection. In some cases, information gained from phasing of alleles in a genome of a subject can be used to tailor a healthcare or therapeutic regimen for the subject. In some cases, one or more methods described herein can be used to construct a database of haplotypes associated with one or more disease or biological trait. A haplotype obtained from a subject using a method as described herein can be compared against a haplotype database, thereby allowing for diagnosis and/or prognosis of a disease or condition. In some cases, haplotype information can be used to aid in the treatment of an infectious disease. For example, bacteria or virus infections can be associated with a plurality of drug resistant markers. Use of one or more phasing methods described herein can help determine if the plurality of drug resistant markers are due to co-infection with single-drug resistant strains or due to infection with a single strain that is multi-drug resistant. One or more phasing methods described herein can also be used to characterize T-cell receptor and/or immunoglobulin repertoire in a subject. For example, one or more phasing methods described herein can be used to assess V(D)J recombination in a subject. One or more phasing methods described herein can be used for diagnosing, predicting, determining or assessing the genetic characteristics of a fetus or embryo. A phasing method described herein can be used to assess genetic phasing within the genome of a developing fetus. In some cases, one or more phasing methods can be used to combat bioterrorism and/or to assess an epidemic or pandemic (e.g., by characterizing or identifying an infectious agent such as a virus or bacterium). In some cases, one or more phasing methods can be used for paternity testing, to establish a family tree, or to determine ancestry. In some cases, one or more phasing methods can be used to determine alternative splicing of RNA transcripts.

EXAMPLES

Example 1 — Characterization of the human oral microbiome by sequential sequencing of bacterial 16S ribosomal operon

[00117] This example describes the characterization of the human oral microbiome by sequencing of the 16S rRNA gene sequences of a number of related bacterial organisms. 16S rRNA gene sequences contain species-specific hypervariable regions that can provide means

IDT Exhibit 1018.0172

for bacterial identification. An exemplary embodiment of 16S rRNA sequential sequencing, described below, is depicted in **FIG. 6**.

Sample nucleic acid

[00118] Microbial genomic DNA is isolated from human saliva using the OMNIgene-DISCOVER sample collection kit (DNA Genotek) according to the manufacturer's instructions. The different line shadings and hash marks indicate different microbial genomes (microbial DNA). The relative numbers (e.g., 4 hatched lines vs 1 gray line) reflect the different ratios of the organisms in the saliva sample. Extracted DNA is then fragmented via sonication to an average length of 1000 bp and purified using Agencourt AMPure XP beads (Beckman Coulter Genomics). In some cases, extracted DNA is not subjected to fragmentation. For example, the microbial genetic DNA can be optionally enriched for target polynucleotides comprising the 16S genomic regions using primers to conserved 16S regions of the genome, thereby generating 16S amplicons.

Generation 16 S libraries with ligated adapters

[00119] The NuGEN Ovation Ultralow Library System (NuGEN Technologies) is used to attach library adaptors to the DNA fragments, thereby generating next generation sequencing libraries from 100 ng of the purified sample according to manufacturer's instructions. Ligation products with a minimum length of 200 bp in length are purified by selective binding to Agencourt AMPure XP beads.

Sequential primer sequencing

[00120] Library members are clonally amplified by bridge amplification on an Illumina flowcell, thereby generating spatially separated clusters, wherein the clusters comprise identical copies of an original sequencing library member. In a first round of sequencing, all clusters, including those containing 16 S ribosomal DNA fragments, are sequenced by Illumina sequencing system using an Illumina standard forward primer. The standard forward primer can be designed to hybridize to a sequencing primer binding site of a library adaptor. Alternatively, a custom primer may be used. The custom primer can be designed to hybridize to a location upstream of and to sequence an informative locus. In such cases, clusters comprising a primer binding site for the first custom primer can be sequenced. The first round of sequencing can generate a first sequencing read ("Read 1"). Following the first sequencing read, the DNA is denatured to wash away the newly synthesized strand comprising Read 1. In cases where the first sequencing primer is a standard forward primer, Read 1 can contain sequence information for the forward end of the cluster templates. In cases where the first sequencing primer is a custom primer, Read 1 can comprise sequence

IDT Exhibit 1018.0173

information for the informative locus. A second round of sequencing can then be performed using a second sequencing primer. The second sequencing primer can be a custom primer that is designed to hybridize to a location upstream of and to sequence an informative locus. In cases where the first round of sequencing is performed using a custom primer that sequences an informative locus, the second sequencing primer is a designed to hybridize to a location upstream of and to sequence informative locus which is downstream of the first informative locus. The second round of sequencing can produce a second sequencing read "Read 2" in clusters comprising a primer binding site for the second sequencing primer. Read 2 can comprise sequence information for the informative locus sequenced by the second sequencing primer. Successive rounds of denaturation, re-priming and sequencing are optionally performed with primers that are designed to sequence additional informative loci. Sequence reads from successive priming and sequencing are compiled for each cluster and aligned to map reads originating from the same nucleic acid fragments.

Example 2 — Genomic DNA sequencing — distinguishing between a gene and its pseudogene using sequential sequencing

[00121] FIG. 7 depicts an exemplary application of a method described herein to distinguish a gene from a pseudogene in a gene/pseudogene pair. Genomic DNA sequencing libraries are prepared using the NuGEN's Encore system. These libraries are sequenced on a DNA sequencing system such as those made by Illumina or Ion Torrent. Clusters include, but are not limited to cluster 1, from a different genomic region from a gene/pseudogene pair, cluster 2, which comprises a portion of the gene of a gene/psuedogene pair, and cluster 3, which comprises a portion of the pseudogene of the gene/pseudogene pair. A first round of sequencing is performed using a first sequencing primer, thereby generating a first sequencing read. In some cases, the first sequencing primer is a standard forward primer as described in Example 1. The first sequencing read reveals that clusters 2 and 3 exhibit different alleles at a certain locus (a T in cluster 2 and a C in cluster 3). However, it is impossible to determine from read 1 which cluster represents the gene or the pseudogene. Following a first sequencing read, the DNA is denatured to wash away the extension strand. An oligonucleotide designed to anneal to a common sequence shared by the gene and pseudogene is injected into the sequencer to act as the second sequencing primer. The second sequencing primer can be designed to hybridize to a conserved sequence shared by the gene and pseudogene, wherein extension of the primer produces a second sequencing read that extends across a locus that can distinguish the gene and pseudogene. An exemplary gene/pseudogene pair which can be distinguished using a method described herein is the

IDT Exhibit 1018.0174

SMN1 gene/SMN2 pseudogene pair. In some embodiments, the second sequencing primer comprises a sequence which is designed to hybridize to a conserved squence shared by SMN1 and SMN2, wherein extension of the primer produces a second sequencing read that extends across a locus that can distinguish SMN1 and SMN2. In one embodiment, the second sequencing primer comprises the sequence CTTCCTTTATTTTCCTTACAGGGT (SEQ. ID. NO: 1). The second sequencing read reveals a downstream C base call at the locus in cluster 3. The C base call reveals that cluster 3 maps to the pseudogene but that cluster 2 maps to the gene. An oligonucleotide set may include primers that will sequence through one of the nucleotide differences between the gene and pseudogene as well as primers that will generate sequence to read nucleotide differences, and therefore determine whether a sequencing read is from a gene or its pseudogene. A combination of such primers will allow multiple gene/pseudogene pairs across the genome to be analyzed simultaneously for genetic mutations.

Example 3 — Targeted DNA sequencing library

[00122] A targeted DNA sequencing library is made using a target-selective library preparation kit from NuGEN, Agilent, Illumina, or Nimblegen, described herein. These libraries are sequenced on a DNA sequencing system such as those made by Illumina or Ion Torrent. Following a first sequencing read, the DNA is denatured to wash away the extension strand. A pool of primers that hybridize to common sequences in gene/pseudogene pairs are injected into the sequencer to act as a priming site for a second sequencing read. A primer set may include primers that will sequence through one of the nucleotide differences between a gene and its pseudogene (e.g., SMN1 and SMN2) as well as primers that will generate sequence to read nucleotide differences, and therefore determine whether a sequencing read is, for example, from a gene or pseudogene. A combination of such primers will allow multiple gene/pseudogene pairs across the genome to be analyzed simultaneously for genetic mutations. This type of technology is useful for genetic diagnostics.

Example 4 — RNA-sequencing library

[00123] An RNA sequencing library is made from NuGEN's Encore Complete RNA-Seq Library System. The library is sequenced on an Illumina DNA sequencer. Following the first sequencing read, a pool of primers that will hybridize to specific exons of interest is injected into the sequencing machine. These primers are used to generate a second sequencing read in a downstream exon. The second, targeted sequencing read provides information about which exons have been spliced together to generate a particular RNA transcript.

IDT Exhibit 1018.0175

[00124] While preferred embodiments of the present invention have been shown and described herein, it will be obvious to those skilled in the art that such embodiments are provided by way of example only. Numerous variations, changes, and substitutions will now occur to those skilled in the art without departing from the invention. It should be understood that various alternatives to the embodiments provided herein described herein may be employed in practicing the invention. It is intended that the following claims define the scope of the invention and that methods and structures within the scope of these claims and their equivalents be covered thereby.

IDT Exhibit 1018.0176

CLAIMS

WHAT IS CLAIMED IS:

1. A method for relating at least two nucleic acid sequences or regions of sequence variation to the same nucleic acid template, the method comprising:

a) creating a nucleic acid library wherein a template is immobilized on a solid support;

b) generating a first sequencing read from the immobilized template with a first oligonucleotide primer, wherein a 3' terminal nucleotide of the first oligonucleotide primer binds to a first binding site in the template;

c) denaturing the first sequencing read from the immobilized template;

d) annealing a second oligonucleotide primer that is complementary to a region or regions within the immobilized template, wherein a 3' terminal nucleotide of the second oligonucleotide primer binds to a second binding site in the template that is more than 5 nt away from the first binding site in the template;

e) generating a second sequencing read with the second oligonucleotide primer; and

f) compiling data from the first sequencing read and the second sequencing read to map reads originating from the template.

2. The method of claim 1, wherein the nucleic acid libraries are generated from amplicons.

3. The method of claim 2, wherein the amplicons comprise conserved regions which flank a variable region.

4. The method of claim 3, wherein alignment of multiple variable regions enables differentiation and/or typing of related transcripts.

5. The method of claim 3, wherein alignment of multiple variable regions enables differentiation and/or typing of related micro-organisms.

6. The method of claim 1, wherein libraries are of reduced complexity.

7. The method of claim 6, wherein reduced complexity is achieved by target enrichment.

8. The method of any of the preceding claims, wherein the nucleic acid library comprises closely related nucleic acid sequences as inserts.

IDT Exhibit 1018.0177

9. A method, comprising:

a) coupling at least one nucleic acid template to a sequencing platform;

b) subjecting the at least one nucleic acid template to a first round of sequencing using a first sequencing primer, wherein the first sequencing primer comprises a 3' terminal nucleotide which anneals to a first binding site in the at least one nucleic acid template, wherein the first round of sequencing produces a first sequencing read of the at least one nucleic acid template;

c) removing the first annealed sequencing primer and the first sequencing read from the at least one nucleic acid template; and

d) subjecting the at least one nucleic acid template to a second round of sequencing using a second sequencing primer, wherein the second sequencing primer comprises a 3' terminal nucleotide which anneals to a second binding site in the at least one nucleic acid template which is more than 5 nt away from the first binding site, wherein the second round of sequencing produces a second sequencing read of the at least one nucleic acid template.

10. The method of claim 9, wherein the first and/or the second round of sequencing comprises sequencing by synthesis.

11. The method of claim 9, wherein the first and/or the second round of sequencing comprises sequencing by ligation.

12. The method of claim 9, wherein the coupling comprises conjugating the at least one nucleic acid template to a solid support of the sequencing platform.

13. The method of claim 12, wherein the solid support comprises a bead, a well, an array, or a flow cell.

14. The method of claim 9, wherein the at least one nucleic acid template comprises an adaptor oligonucleotide ligated to one or both ends of the at least one nucleic acid template.

15. The method of claim 14, wherein the first sequencing primer is selectively hybridizable to a primer binding site of the adaptor oligonucleotide.

16. The method of claim 14, wherein the second sequencing primer is not selectively hybridizable to the primer binding site of the adaptor oligonucleotide.

17. The method of claim 9, wherein the second sequencing read comprises an informative locus.

18. The method of claim 17, wherein the informative locus is a polymorphic locus.

IDT Exhibit 1018.0178

19. The method of claim 17, wherein the first sequencing read comprises a first informative locus and wherein the second sequencing read comprises a second informative locus.

20. The method of claim 19, comprising determining that a first allele detected at the first informative locus and a second allele detected at the second informative locus are in phase.

21. The method of claim 17, wherein the second sequencing primer is designed to hybridize to a location that is within *x* nt upstream of the informative locus, wherein *x* is a number that is less than an average sequencing read length produced by the first or second round of sequencing.

22. The method of claim 9, wherein the removing comprises a denaturing step.

23. The method of claim 22, wherein the removing further comprises a washing step.

24. The method of claim 9, further comprising subjecting the at least one nucleic acid template to one or more additional rounds of sequencing, wherein the one or more additional rounds of sequencing comprise the steps of : (i) removing a previously annealed sequencing primer and previously generated sequencing read from the at least one nucleic acid template; (ii) annealing an additional sequencing primer to the at least one nucleic acid template; and (iii) subjecting the at least one nucleic acid template to an additional round of sequencing using the additional sequencing primer, thereby producing an additional sequencing read of the at least one nucleic acid template.

25. The method of claim 24, wherein the additional sequencing primer comprises a 3' terminal nucleotide which hybridizes to a binding site that is more than 5 nt away from a binding site of the 3' terminal nucleotide of the previously annealed sequencing primer.

26. The method of claim 24, wherein a first allele detected by the first sequencing read, a second allele detected by the second sequencing read, and an additional allele detected by the additional sequencing read are determined to be in phase.

27. The method of claim 24, wherein the additional sequencing primer and the previously annealed sequencing primer are designed to hybridize to genomic locations that map to a single chromosome.

28. The method of claim 27, wherein the genomic locations that map to a single chromosome span a distance that is less than a length of the at least one nucleic acid template.

29. The method of claim 9, wherein the at least one nucleic acid template is a member of a sequencing library.

IDT Exhibit 1018.0179

30. The method of claim 29, wherein the sequencing library is a whole-genome sequencing library.

31. The method of claim 29, wherein the sequencing library is a target-enriched sequencing library.

32. The method of claim 29, wherein the sequencing library comprises nucleic acid fragments of an average length.

33. The method of claim 32, wherein the average length is about 50 nucleotides to about 10 kb.

34. The method of any of claims 9-33, wherein the first, second, and optionally additional sequencing reads comprise sequence information from a plurality of loci.

35. The method of claim 34, wherein phasing of alleles detected at the plurality of loci is determined based upon the first, second, and optionally additional sequencing reads.

36. The method of any of claims 9-33, wherein the at least one nucleic acid template comprises a plurality of nucleic acid templates.

37. The method of any of claims 9-33, wherein the at least one nucleic acid template comprises at least 10,000 nucleic acid templates.

38. The method of claim 37, wherein the at least one nucleic acid template comprises 3 billion nucleic acid templates.

39. The method of any of claims 9-33, wherein the at least one nucleic acid template is coupled to a unique addressable location of the sequencing platform.

40. The method of claim 39, comprising compiling sequencing reads from the unique addressable location into one data file.

41. The method of claim 39, comprising determining phasing of a plurality of alleles detected across two or more unique addressable locations.

42. The method of claim 41, wherein the two or more unique addressable locations each comprise two or more detected alleles, wherein the two or more detected alleles of each of the two or more unique addressable locations are determined to be in phase if at least one common detected allele is shared by the two or more unique addressable locations.

43. The method of any of claims 9-33, wherein step (d) comprises subjecting the at least one nucleic acid template to a second round of sequencing using a mixture of second sequencing primers.

IDT Exhibit 1018.0180

44. The method of claim 43, wherein the mixture of second sequencing primers comprises a pair of primers, wherein each primer in the primer pair is designed to sequence a forward or reverse complement strand containing a locus.

45. The method of claim 43, wherein primers within the mixture are designed and/or selected to have minimal potential to hybridize to the same nucleic acid template.

46. A kit, comprising:

a) one or more oligonucleotide primers for performing a sequencing reaction; and

b) instructions for practicing a method of any of the preceding claims using the one or more oligonucleotide primers.

47. A computer readable medium comprising computer executable code that, upon execution by a computer processor, implements a method comprising:

a) accessing a data file in computer memory that comprises a plurality of sequencing reads corresponding to a first addressable location of a sequencing platform, wherein the sequencing reads are generating by subjecting a first nucleic acid template coupled to the addressable location to a first and at least a second round of sequencing, wherein the first and at least second rounds of sequencing are performed using a first and second sequencing primer, respectively;

b) using the data file, detecting a first detected allele at a first locus of the nucleic acid template and detecting a second detected allele at a second locus of the nucleic acid template; and

c) determining the first and second detected allele to be in phase based upon the detecting.

48. The computer readable medium of claim 47, wherein the method further comprises:

d) accessing a second data file in computer memory that comprises a plurality of sequencing reads corresponding to a second addressable location of a sequencing platform, wherein the sequencing reads are generated by subjecting a second nucleic acid template coupled to the addressable location to the first and the at least second round of sequencing;

e) using the second data file, detecting a first detected allele at a first locus of the second nucleic acid template and detecting a second detected allele at a second locus of the second nucleic acid template; and

IDT Exhibit 1018.0181

f) determining the first and second detected alleles of the second nucleic acid template to be in phase with the first and second detected alleles of the first nucleic acid template if the first or second detected allele of the second nucleic acid template is shared in common with the first nucleic acid template.

IDT Exhibit 1018.0182

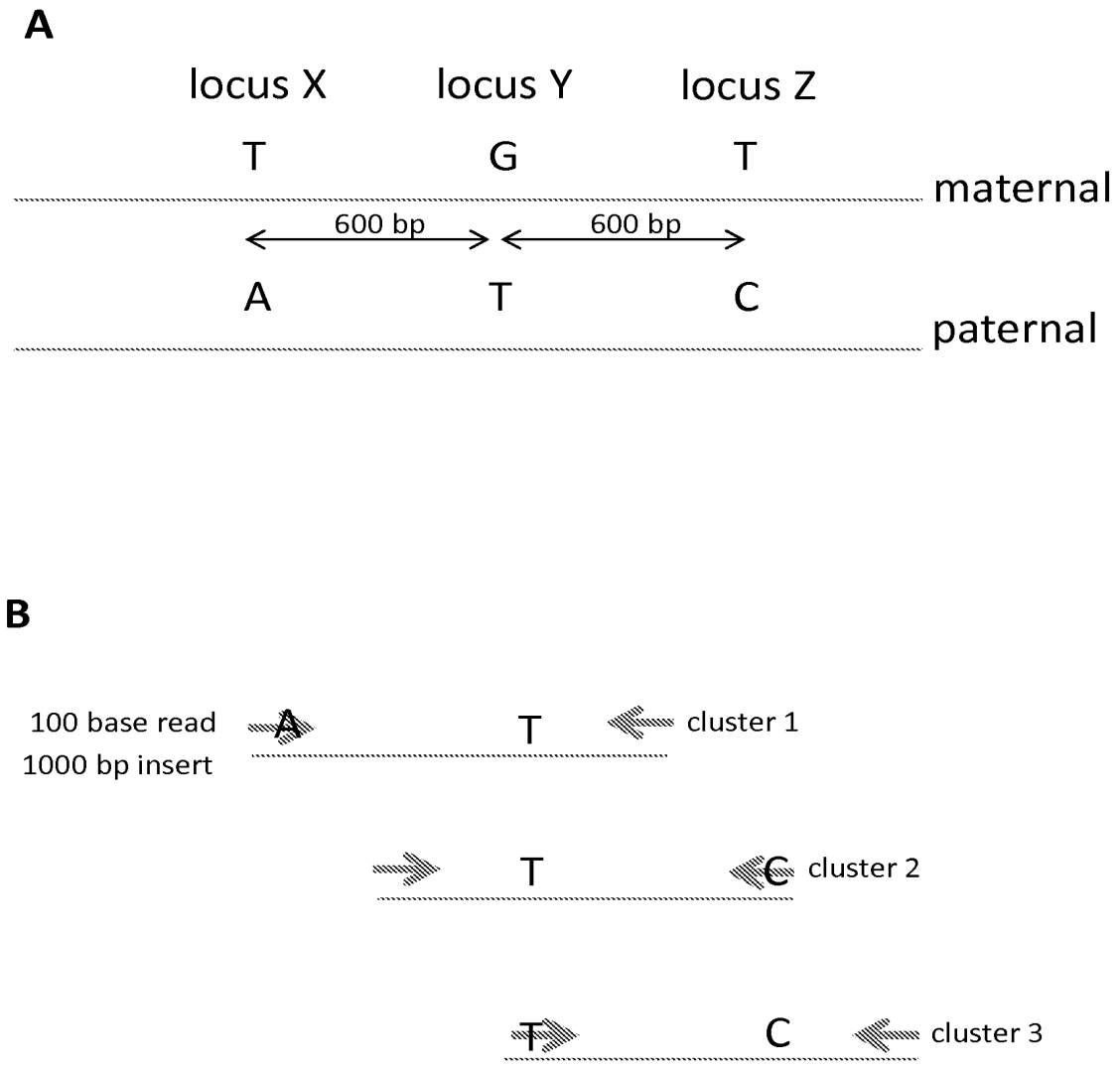

A
locus X
locus Y
locus Z
T
G
T
maternal
600 bp
600 bp
A
T
C
paternal
B
100 base read
1000 bp insert
A
T
cluster 1
T
C
cluster 2
T
C
cluster 3

FIG. 1

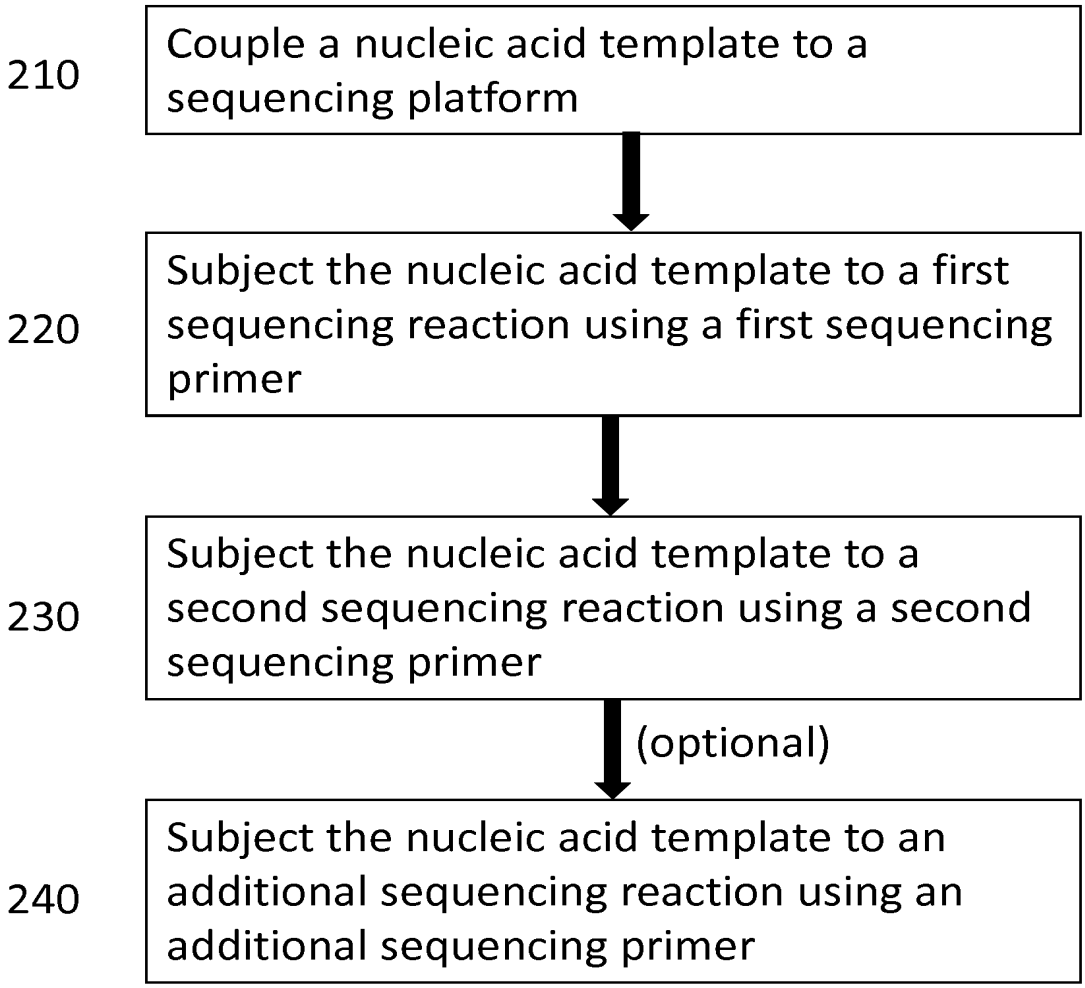

210
Couple a nucleic acid template to a sequencing platform
220
Subject the nucleic acid template to a first sequencing reaction using a first sequencing primer
230
Subject the nucleic acid template to a second sequencing reaction using a second sequencing primer
(optional)
240
Subject the nucleic acid template to an additional sequencing reaction using an additional sequencing primer

FIG. 2

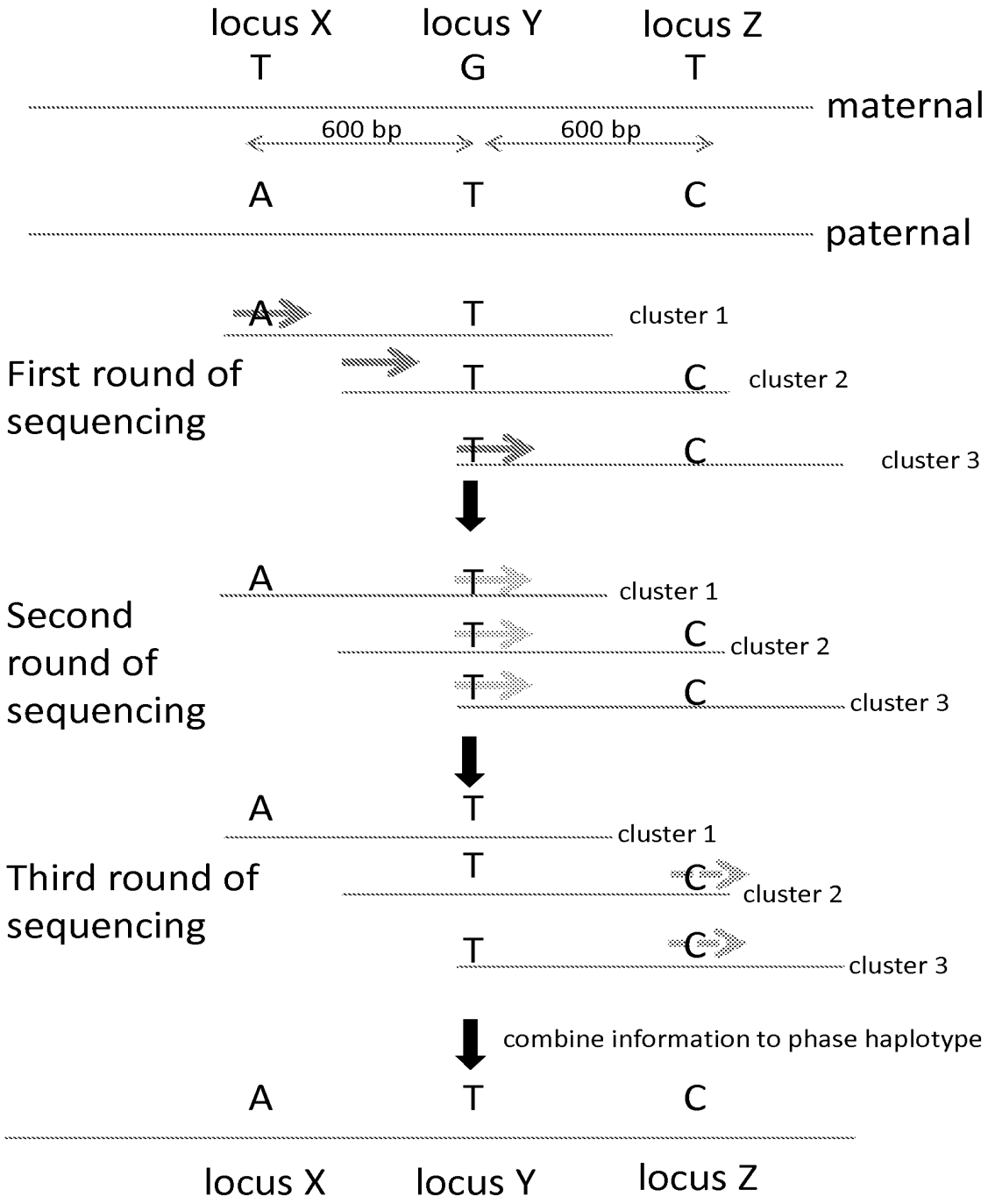

locus X
locus Y
locus Z
T
G
T
maternal
600 bp
600 bp
A
T
C
paternal
First round of sequencing
A
T
cluster 1
T
C
cluster 2
T
C
cluster 3
Second round of sequencing
A
T
cluster 1
T
C
cluster 2
T
C
cluster 3
Third round of sequencing
A
T
cluster 1
T
C
cluster 2
T
C
cluster 3
combine information to phase haplotype
A
T
C
locus X
locus Y
locus Z

FIG. 3

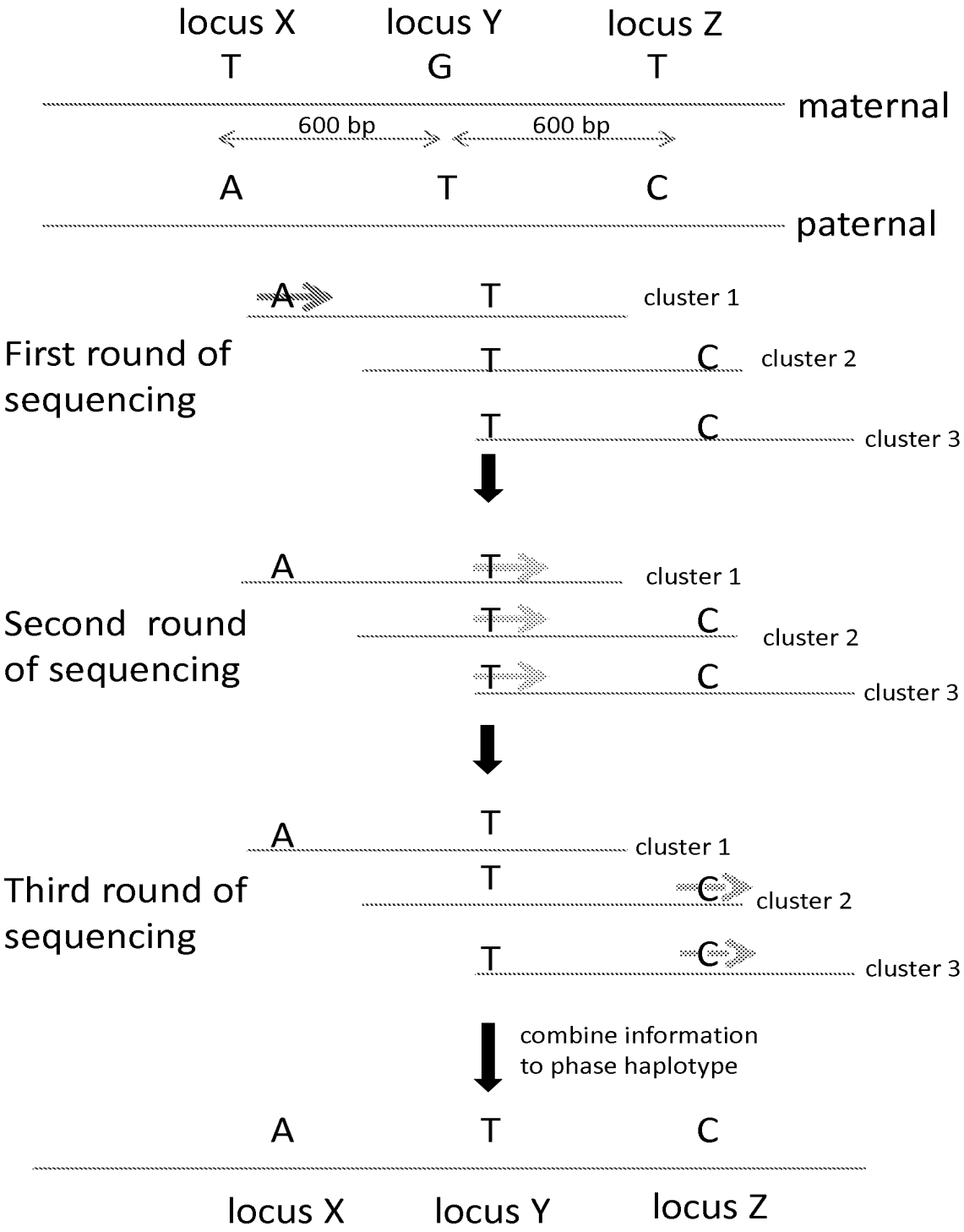

locus X
locus Y
locus Z
T
G
T
maternal
600 bp
600 bp
A
T
C
paternal
First round of sequencing
cluster 1
cluster 2
cluster 3
Second round of sequencing
cluster 1
cluster 2
cluster 3
Third round of sequencing
cluster 1
cluster 2
cluster 3
combine information to phase haplotype
A
T
C
locus X
locus Y
locus Z

FIG. 4

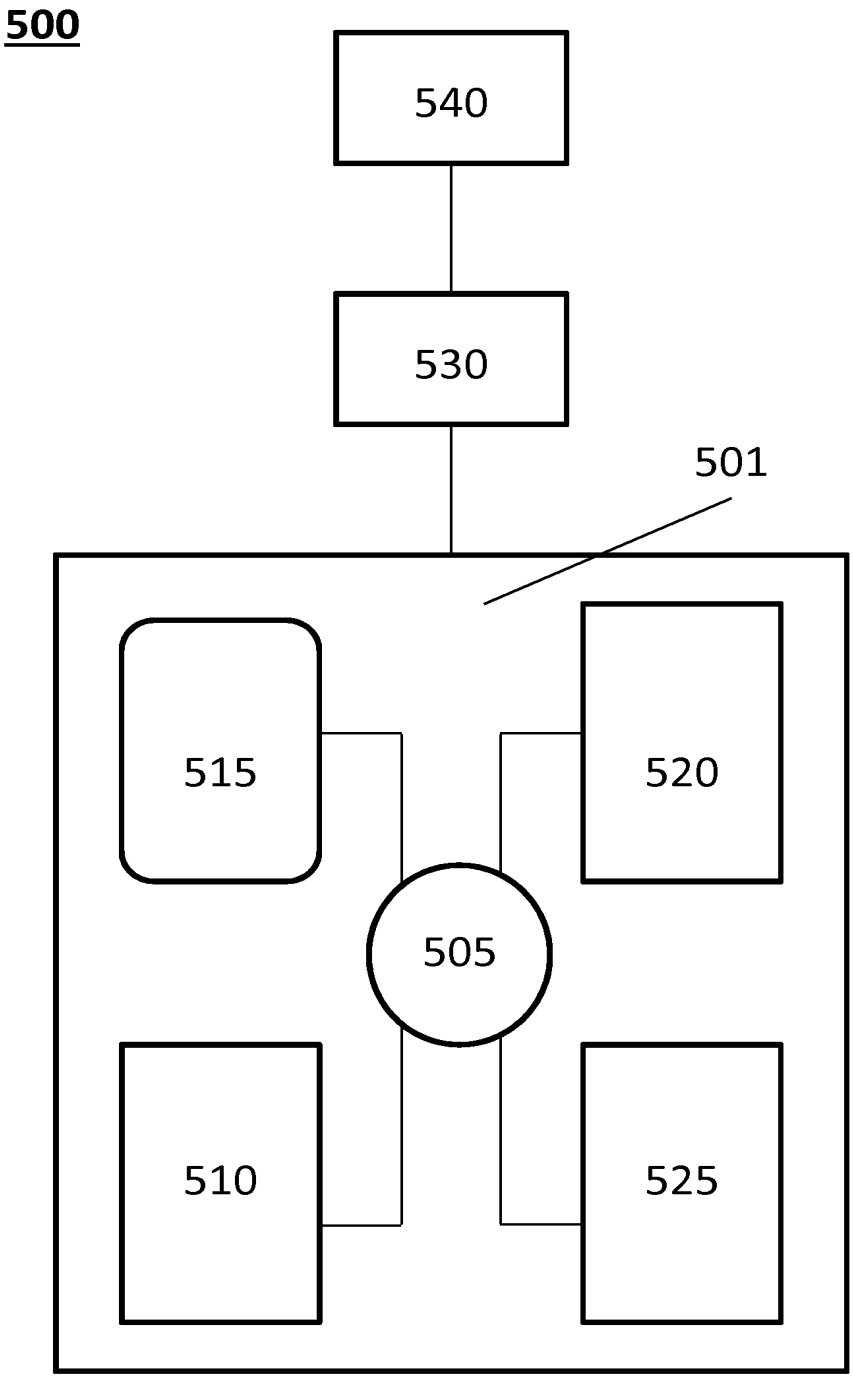


FIG. 5

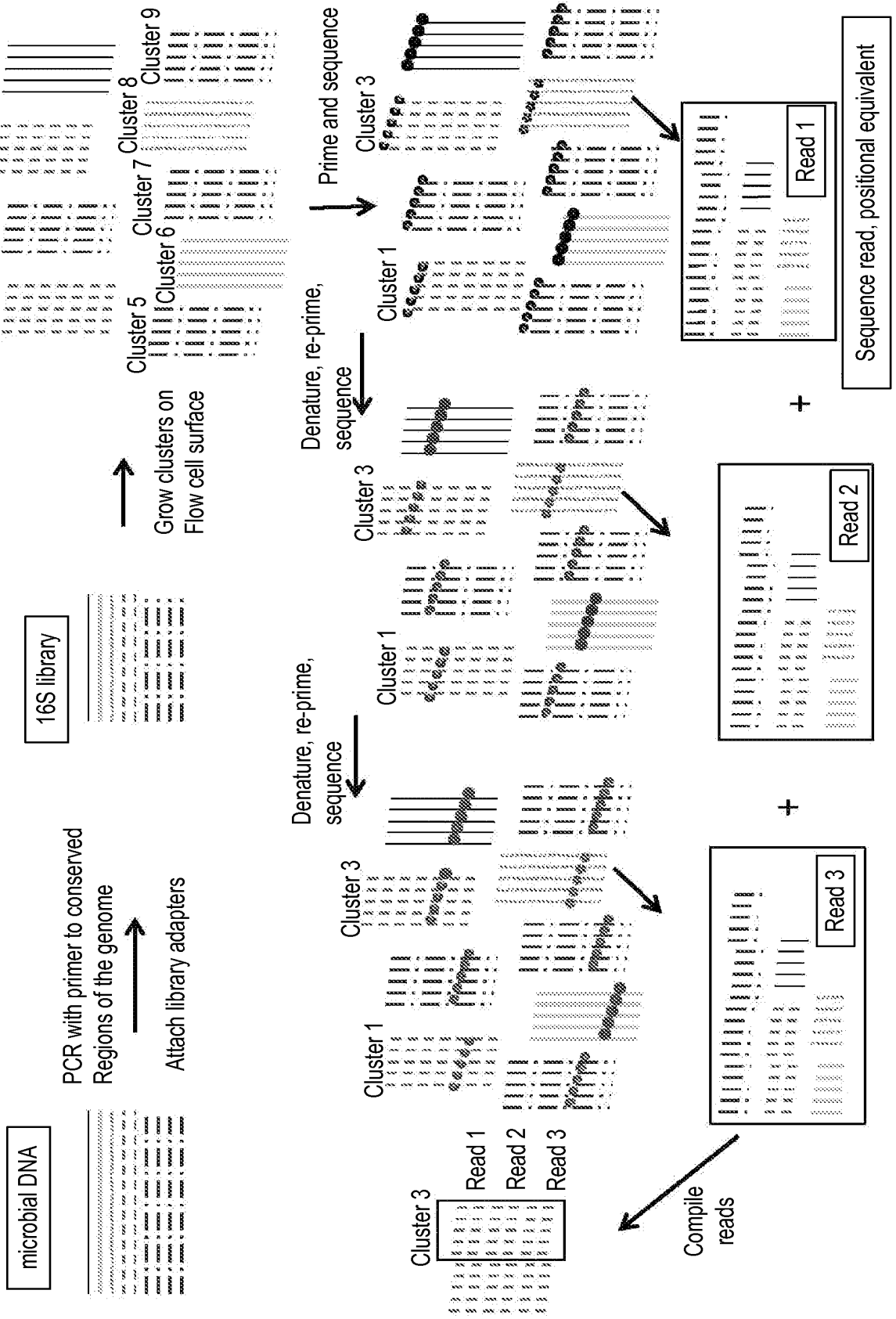

microbial DNA
PCR with primer to conserved Regions of the genome
Attach library adapters
16S library
Grow clusters on Flow cell surface
Cluster 1
Cluster 2
Cluster 3
Cluster 4
Cluster 5
Cluster 6
Cluster 7
Cluster 8
Cluster 9
Denature, re-prime, sequence
Prime and sequence
Read 1
Read 2
Read 3
Compile reads
Sequence read, positional equivalent

FIG. 6

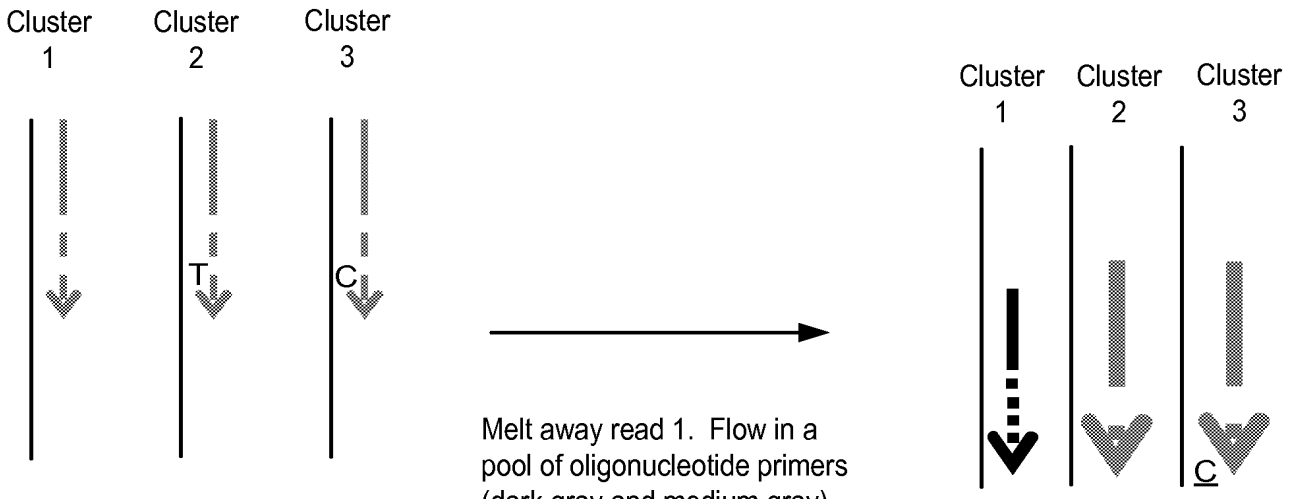


FIG. 7

IDT Exhibit 1018.0189

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US2014/028356

A. CLASSIFICATION OF SUBJECT MATTER
IPC(8) - C12Q 1/68 (2014.01)
USPC - 435/6.11
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(8) - C12Q 1/68; C40B 20/02, 20/04, 20/06, 20/08, 30/00, 30/10, 40/06, 50/06 (2014.01)
USPC - 435/6.11, 6.12, 6.15; 506/4, 5, 6, 7, 16, 26; 536/23.1

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched
CPC - A61K 38/00; B01J 19/0046, 2219/00585, 2219/00596, 2219/00707, 2219/00722; C12Q 1/6827, 1/6883, 1/689 (2014.06)

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)
Orbit, Google Patents, Google Scholar

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X --- Y	US 2009/0298075 A1 (TRAVERS et al) 03 December 2009 (03.12.2009) entire document	1-36, 39-45, 47, 48 --- 37, 38
Y	US 2010/0105052 A1 (DRMANAC et al) 29 April 2010 (29.04.2010) entire document	37, 38
A	US 2012/0071331 A1 (CASBON et al) 22 March 2012 (22.03.2012) entire document	1-45, 47, 48
A	US 2011/0319290 A1 (RAYMOND et al) 29 December 2011 (29.12.2011) entire document	1-45, 47, 48
A	US 2011/0105364 A1 (KURN et al) 05 May 2011 (05.05.2011) entire document	1-45, 47, 48

☐ Further documents are listed in the continuation of Box C. ☐

* 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
26 June 2014	15 JUL 2014
Name and mailing address of the ISA/US Mail Stop PCT, Attn: ISA/US, Commissioner for Patents P.O. Box 1450, Alexandria, Virginia 22313-1450 Facsimile No. 571-273-3201	Authorized officer: Blaine R. Copenheaver PCT Helpdesk: 571-272-4300 PCT OSP: 571-272-7774

Form PCT/ISA/210 (second sheet) (July 2009)

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US2014/028356

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.:
because they relate to subject matter not required to be searched by this Authority, namely:

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.: 46
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.

IDT Exhibit 1018.0191

PTO/SB/08a (07-09)
Approved for use through 07/31/2012. OMB 0651-0031
U.S. Patent and Trademark Office; U.S. DEPARTMENT OF COMMERCE
Under the Paperwork Reduction Apt of 1995 no Persons are required to respond to a collection of information unless it contains a valid OMR control number

Substitute for form 1449A/PTO

INFORMATION DISCLOSURE STATEMENT BY APPLICANT

(Use as many sheets as necessary)

Sheet 1 of 1

Complete if Known	
Application Number	14/540,917
Filing Date	November 13, 2014
First Named Inventor	AMORESE, Douglas
Art Unit	1637
Examiner Name	STRZELECKA, Teresa E.
Attorney Docket Number	NUGN-001/01US 301735-2003

U. S. PATENT DOCUMENTS

Examiner Initials*	Cite No.[1]	Document Number Number-Kind Code[2] (if known)	Publication Date MM-DD-YYYY	Name of Patentee or Applicant of Cited Document	Pages, Columns, Lines, Where Relevant Passages or Relevant Figures Appear
	1.	2012/0208705 A1	08-16-2012	STEEMERS et al.	

FOREIGN PATENT DOCUMENTS

Examiner Initials*	Cite No.[1]	Foreign Patent Document Country Code[3] "Number[4] "Kind Code[5] *(if known)*	Publication Date MM-DD-YYYY	Name of Patentee or Applicant of Cited Document	Pages, Columns, Lines, Where Relevant Passages or Relevant Figures Appear	T[6]
	2.	WO 2008/093098 A2	08-07-2008	SOLEXA LIMITED		
	3.	WO 2012/061832 A1	05-10-2012	ILLUMINA, INC.		
	4.	WO 2014/144092 A1	09-18-2014	NUGEN TECHNOLOGIES, INC.		

NON PATENT LITERATURE DOCUMENTS

Examiner Initials*	Cite No.[1]	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[2]
	5.	MYERS, R., "Protocol for Creating Multiplexed miRNA Libraries for Use in Illumina Sequencing", *Myers lab microRNA-seq Protocol, Hudson Alpha Institute for Biotechnology,* May 2, 2013, pages 1-15. Retrieved from the Internet:<https://www.encodeproject.org/documents/9fb8e06c-d145-447a-a606-e7f3dc44b549/@@download/attachment/HA_Myers_miRNA-seq_Protocol_2013-05-02.pdf> on February 28, 2015. entire document	
	6.	International Search Report and Written Opinion in international Application No. PCT/US2014/065530, mailed March 5, 2015, 14 pages.	

Examiner Signature		Date Considered	

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

[1] Applicant's unique citation designation number (optional). [2] Applicant is to place a check mark here if English language Translation is attached.

This collection of information is required by 37 CFR 1.98. 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 2 hours 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, 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.

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

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
7 August 2008 (07.08.2008)

PCT

(10) International Publication Number
WO 2008/093098 A2

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

(21) International Application Number:
PCT/GB2008/000337

(22) International Filing Date: 1 February 2008 (01.02.2008)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
60/899,221 2 February 2007 (02.02.2007) US

(71) Applicant *(for all designated States except US)*: **SOLEXA LIMITED** [GB/GB]; Chesterford Research Park, Little Chesterford, Saffron Walden Essex CB10 1XL (GB).

(72) Inventors; and

(75) Inventors/Applicants *(for US only)*: **BIGNALL, Helen** [GB/GB]; SOLEXA LIMITED, Chesterford Research Park, Little Chesterford, Saffron Walden Essex CB10 1XL (GB). **FRASER, Louise** [GB/GB]; SOLEXA LIMITED, Chesterford Research Park, Little Chesterford, Saffron Walden Essex CB10 1XL (GB). **GORMLEY, Niall** [IE/GB]; SOLEXA LIMITED, Chesterford Research Park, Little Chesterford, Saffron Walden Essex CB10 1XL (GB).

(74) Agents: **WHITE, Nina, Louise** et al.; BOULT WADE TENNANT, Verulam Gardens, 70 Gray's Inn Road, London WC1X 8BT (GB).

(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, 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, PG, PH, PL, PT, RO, RS, RU, SC, SD, SE, SG, SK, SL, SM, 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, MT, NL, NO, PL, PT, RO, SE, SI, SK, 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*

— *with sequence listing part of description published separately in electronic form and available upon request from the International Bureau*

WO 2008/093098 A2

(54) Title: METHODS FOR INDEXING SAMPLES AND SEQUENCING MULTIPLE NUCLEOTIDE TEMPLATES

(57) Abstract: The invention relates to methods for indexing samples during the sequencing of polynucleotide templates, resulting in the attachment of tags specific to the source of each nucleic acid sample such that after a sequencing run, both the source and sequence of each polynucleotide can be determined. Thus, the present invention pertains to analysis of complex genomes (e.g., human genomes), as well as multiplexing less complex genomes, such as those of bacteria, viruses, mitochondria, and the like.

IDT Exhibit 1018.0193

Methods for indexing samples and sequencing multiple polynucleotide templates

Field of the invention

The invention relates to methods for indexing samples during the sequencing of polynucleotide templates, resulting in the attachment of tags specific to the source of each nucleic acid sample such that after a sequencing run, both the source and sequence of each polynucleotide can be determined. Thus, the present invention pertains to analysis of complex genomes (e.g., human genomes), as well as multiplexing less complex genomes, such as those of bacteria, viruses, mitochondria, and the like.

Background to the invention

Several publications and patent documents are referenced in this application in order to more fully describe the state of the art to which this invention pertains. The disclosure of each of these publications and documents is incorporated by reference herein.

Advances in the study of biological molecules have been led, in part, by improvement in technologies used to characterise the molecules or their biological reactions. In particular, the study of the nucleic acids DNA and RNA has benefited from developing technologies used for sequence analysis.

The study of complex genomes, in particular, the search for the genetic basis of disease in humans requires genotyping on a massive scale. Screens for numerous genetic markers performed on populations large enough to yield statistically significant data are needed before associations can be made between a given genotype and a

particular disease. However large-scale genotyping is demanding in terms of the cost of both materials and labour involved, and the time taken to perform the study, especially if the methodology employed involves separate serial analysis of individual DNA samples. One shortcut is to pool DNA from many individuals and to determine parameters such as the ratio of changes at certain positions in the genome. Such measurements of 'allele frequency' in the pool of samples can be used to correlate the relationship between the changes in the genome sequence and the occurrence of a disease. Hence, an association study involving 1000 patients would in theory only necessitate a 'one-pot' reaction for each genetic change. Pooling therefore represents an effective technique for analysing large quantities of samples in a facile manner.

One disadvantage of pooling samples prior to analysis is that information pertaining to individual DNA samples is lost; only global information such as allele frequencies is gathered, as there is no easy method for discerning which individuals gave rise to a particular genotype. An ability to genotype large populations in a small number of reactions, while retaining the information relating to the source of the individual samples, would yield the information content of a full non-pooled population screen in the time and at the cost of a pooled reaction.

Several of the new methods employed for high throughput DNA sequencing (*Nature*. 437, 376-380 (2005); *Science*. 309, 5741, 1728-1732 (2005)) rely on a universal amplification reaction, whereby a DNA sample is randomly fragmented, then treated such that the ends of the different fragments all contain the same DNA sequence. Fragments with universal ends can be amplified in a single reaction with a single pair of

amplification primers. Separation of the library of fragments to the single molecule level prior to amplification ensures that the amplified molecules form discrete populations that can then be further analysed. Such separations can be performed either in emulsions (*Nature*. 437, 376-380 (2005); *Science*. 309, 5741, 1728-1732 (2005)), or on a surface (Nucleic Acids Research 27, e34 (1999); Nucleic Acids Research 15, e87 (2000)).

WO 98/44151 and WO 00/18957 both describe methods of forming polynucleotide arrays based on 'solid-phase' nucleic acid amplification, which is a bridging amplification reaction wherein the amplification products are immobilised on a solid support in order to form arrays comprised of nucleic acid clusters or 'colonies'. Each cluster or colony on such an array is formed from a plurality of identical immobilised polynucleotide strands and a plurality of identical immobilised complementary polynucleotide strands. The arrays so-formed are generally referred to herein as 'clustered arrays' and their general features will be further understood by reference to WO 98/44151 or WO 00/18957, the contents of both documents being incorporated herein in their entirety by reference.

In common with all amplification techniques, solid-phase bridging amplification requires the use of forward and reverse amplification primers which include 'template-specific' nucleotide sequences which are capable of annealing to sequences in the template to be amplified, or the complement thereof, under the conditions of the annealing steps of the amplification reaction. The sequences in the template to which the primers anneal under conditions of the amplification reaction may be referred to herein as 'primer-binding' sequences.

Certain embodiments of the methods described in WO 98/44151 and WO 00/18957 make use of 'universal' primers to amplify templates comprising a variable template portion that it is desired to amplify flanked 5' and 3' by common or 'universal' primer binding sequences. The 'universal' forward and reverse primers include sequences capable of annealing to the 'universal' primer binding sequences in the template construct. The variable template portion, or 'target' may itself be of known, unknown or partially known sequence. This approach has the advantage that it is not necessary to design a specific pair of primers for each target sequence to be amplified; the same primers can be used for amplification of different templates provided that each template is modified by addition of the same universal primer-binding sequences to its 5' and 3' ends. The variable target sequence can therefore be any DNA fragment of interest. An analogous approach can be used to amplify a mixture of templates (targets with known ends), such as a plurality or library of target nucleic acid molecules (e.g. genomic DNA fragments), using a single pair of universal forward and reverse primers, provided that each template molecule in the mixture is modified by the addition of the same universal primer-binding sequences.

DNA from more than one source can be sequenced on an array if each DNA sample is first tagged to enable its identification after it has been sequenced. Many low scale DNA-tag methodologies already exist, for example fluorescent labelling (Haughland, Handbook of Fluorescent Probes and Research Products, Invitrogen/Molecular Probes), but these are limited in scope to less than 10 or so reactions in parallel. DNA tags can be added to the ends of DNA fragments by cloning, of example as described in US 5604097. The tags

consist of eight four base 'words', where each word uses only three bases (A,T and C) in various combinations resulting in a total of 16,777,216 different tags that all have the same base pair composition and melting points. Such tags are used to label target molecules in a sample so that after an amplification reaction, each original molecule in the sample has a unique tag. The tags can then be used to 'sort' the sample onto beads containing sequences complementary to the tags such that each bead contains multiple copies of a single amplified target sequence (Brenner et al., (2000) Nature Biotechnology, 18, 630). In this application the tags are not sequenced, so the method does not provide a method of analysing targets from multiple samples, but rather a method of sorting a mixture of amplified templates from a single sample. The problem with enabling the method for individual samples rather than individual molecules is that the tags are synthesised in a combinatorial manner, meaning that all 16,777,216 different sequences are obtained in a single mixture in the same tube. Whilst this is ideal for treating one sample such that each individual molecule in the sample carries a different tag, it does not permit attachment of the same tag to every molecule in the sample.

DNA samples from multiple sources can, however, be tagged with different nucleic acid tags such that the source of the sample can be identified. Previous application WO05068656 describes the generic concept of indexing samples. In order to utilise this invention on arrays of amplified single molecule templates, for example as described in WO9844151, WO06099579 or WO04069849, it is advantageous to prepare the nucleic acids using the novel method described herein. The optimised DNA sample

preparation techniques described herein are applicable to any method where the samples are amplified prior to sequencing. The DNA sample preparation techniques presented herein describe in detail the optimal placements of the sequencing primers and indexed tags within the DNA constructs to be sequenced.

Summary of the invention

The present inventors have developed methods for indexing samples, wherein the samples are amplified from isolated single template molecules. Using the techniques of the invention, it is possible to prepare a nucleic acid sample for a sequencing reaction wherein both the target sequence and the tag sequence can be determined.

In a first embodiment of the invention, a method is presented for sequencing nucleic acid sequences on an array and identifying subsets of nucleic acid sequences on an array, wherein each subset of nucleic acid sequences is isolated from a different source, wherein the method comprises the steps of:

(a) providing at least two samples of randomly fragmented double stranded nucleic acid targets, wherein each of said randomly fragmented double stranded nucleic acid targets is isolated from a different source;

(b) ligating a universal adaptor to the ends of each target fragment of each sample to generate adaptor-target-adaptors of each sample, wherein each of said adaptor-target-adaptors comprises a target fragment flanked by universal adaptor sequences and said universal adaptor comprises a region of double stranded nucleic acid and at least one region of single stranded nucleic acid;

(c) amplifying adaptor-target-adaptors of each sample with two or more sample specific amplification primers to generate amplified nucleic acids, wherein one of said amplification primers comprises a sample specific tag sequence, and wherein amplified nucleic acids of each sample comprise said sample specific tag sequence and said sample specific tag sequence differentiates amplified nucleic acids originating from different samples;

(d) pooling the amplified nucleic acids of different samples;

(e) immobilising the amplified nucleic acids of different samples on an array to generate an array of immobilised fragments;

(f) sequencing the immobilised fragments on the array to determine a sequence read of each immobilised target fragment and identify the sample specific tag sequence of each immobilised fragment, thereby determining both a nucleic acid sequence of said immobilized fragment and identifying the immobilized fragment as a member of a subset of nucleic acids on the array.

In a second embodiment of the invention, a method is presented for sequencing nucleic acid sequences on an array and identifying subsets of nucleic acid sequences on an array, wherein each subset of nucleic acid sequences is isolated from a different source, wherein the method comprises the steps of:

(a) providing at least two samples of randomly fragmented double stranded nucleic acid targets, wherein each of said randomly fragmented double stranded nucleic acid targets is isolated from a different source;

(b) ligating a sample specific tagged adaptor to the ends of each target fragment of each sample to generate adaptor-

target-adaptors of each sample, wherein each of said adaptor-target-adaptors comprises a target fragment flanked by tagged adaptor sequences and said tagged adaptor comprises a region of double stranded nucleic acid and at least one region of single stranded nucleic acid and a sample specific tag that differentiates adaptor-target-adaptors originating from different samples;

(c) pooling the adaptor-target-adaptors of different samples to generate a pooled sample of adaptor-target-adaptors;

(d) amplifying the pooled sample of adaptor-target-adaptors with a pair of universal amplification primers to generate an amplified pooled sample;

(e) immobilising the amplified pooled sample on an array to generate an array of immobilised fragments;

(f) sequencing the array to determine a sequence read of each immobilised target fragment and identify the sample specific tag sequence of each immobilised fragment, thereby determining both a nucleic acid sequence of said immobilized fragment and identifying the immobilized fragment as a member of a subset of nucleic acids on the array.

In a third embodiment of the invention, a method is presented for sequencing nucleic acid sequences on an array and identifying subsets of nucleic acid sequences on an array, wherein each subset of nucleic acid sequences is isolated from a different source, wherein the method comprises the steps of:

(a) providing at least two samples of randomly fragmented double stranded nucleic acid targets, wherein each of said randomly fragmented double stranded nucleic acid targets is isolated from a different source;

(b) ligating a sample specific tagged adaptor to the ends of each target fragment of each sample to generate adaptor-target-adaptors of each sample, wherein each of said adaptor-target-adaptors comprises a target fragment flanked by tagged adaptor sequences and said tagged adaptor comprises a region of double stranded nucleic acid and at least one region of single stranded nucleic acid and a sample specific tag that differentiates adaptor-target-adaptors originating from different samples, and wherein the sample specific tag sequence is attached directly to the target fragment in each adaptor-target-adaptor with no intervening nucleotides between the sample specific tag sequence and target fragment sequences;

(c) amplifying adaptor-target-adaptors of each sample with a pair of amplification primers complementary to the sample specific tag sequence to generate amplified nucleic acids from each sample;

(d) pooling the amplified nucleic acids from each sample to generate a pool of amplified nucleic acid samples;

(e) immobilising the pool of amplified nucleic acid samples on an array to generate an array of immobilised fragments;

(f) sequencing the array to determine a sequence read of each immobilised target fragment and identify the sample specific tag sequence of each immobilised fragment, thereby determining both a nucleic acid sequence of said immobilized fragment and identifying the immobilized fragment as a member of a subset of nucleic acids on the array.

In accordance with the first or second embodiment of the invention, sequencing can be performed in two separate reads. In accordance with the third embodiment of the invention, sequencing of the target fragment and the sample

specific tag sequence can be performed in a single sequencing read.

With respect to the first, second, or third embodiments, the method encompasses a step wherein the pooled samples are amplified on a solid surface prior to sequencing. In a particular aspect, the solid surface on which amplification is performed is a collection of beads. In a more particular aspect, the collection of beads may be isolated as single beads in an emulsion.

With respect to the first, second, or third embodiments, the method encompasses a step wherein the array of immobilized fragments is amplified prior to sequencing to generate an amplified array. In a particular aspect, an amplified array so produced is a clustered array of amplified single molecules. Such clustered arrays may be formed, for example, by solid-phase nucleic acid amplification or isothermal solid-phase nucleic acid amplification.

The present method encompasses sequencing that involves cycles of ligation with labelled oligonucleotides or cycles of addition of nucleotides. In a particular aspect of the invention, the nucleotides are labelled. In a more particular aspect the nucleotides are labelled with fluorophores.

In an aspect of any of the first, second, or third embodiments of the invention, the sequencing step may be carried out on an array of immobilised beads.

A further aspect of the invention comprises a kit comprising a double stranded nucleic acid adapter and at least two different primer polynucleotides that carry a different tag sequence that does not hybridise to the adapter.

Brief description of the figures

Figure 1 shows a schematic of an indexed sample construct using the first method of the invention, using universal adaptors. Common adaptors are used for all tags, tags are added using one unique amplification primer and sequencing is carried out in two reads.

Figure 2 shows a schematic of an indexed sample construct using the second method of the invention, using tagged adaptors. Tagged adaptors are used for each sample and amplified with one pair of universal primers. Sequencing is carried out in two reads.

Figure 3 shows a schematic of an indexed sample construct using the third method of the invention, using adjacent tag-targets. Two adaptors and two amplification primers are used per tag and sequencing carried out in a single read.

Figure 4 shows numerical data on two sequencing reads obtained from a sample prepared according to the first method of the invention, amplified into an array of clusters and sequenced with two sequencing reads, one for the target and one for the tag.

Figure 5 shows the concept of tagging a sample for use in a paired end reading process. Libraries are amplified at the PCR step using a nested PCR reaction with 3 primers (2 universal and 1 tag oligo.

Figure 6a shows a nested PCR reaction using three amplification primers, one of which carries the tag.

Figure 6b shows the same nested PCR concept as figure 6, but with the benefit of the sequences of the various adapters and primers.

Figure 7 shows the embodiment of the method where the first sequencing read is performed, followed by removing the first sequencing primer, hybridising a second sequencing primer to determine the tag sequence, inverting the strand on the surface using a strand resynthesis protocol, and performing a second read at the opposite end of the fragment to the first read to generate a pair of reads from each fragment along with the tag read.

Figure 8 shows some exemplary data obtained using the method shown in figure 7.

Figure 9 shows the sequence of the various oligonucleotides used in the process

Figure 10 shows a schematic of MJ Research thermocycler, coupled with an 8-way peristaltic pump Ismatec IPC ISM931 equipped with Ismatec tubing (orange/yellow, 0.51 mm ID).

Detailed description of the invention

The present invention is directed to methods that advance techniques for nucleic acid analysis, such that target nucleic acids from a plurality of sources can be tagged with a unique, identifiable sequence of nucleic acid bases and then sequenced on a single array. As described

herein, the presence of a characteristic nucleic acid sequence tag on an immobilised target molecule permits identification of the source of the target, concurrent with a sequence read from the target. This is a dramatic improvement over pre-existing array technologies which generally require a two-step process involving an initial sequencing step for sequencing the pooled nucleic acid targets, followed by a second analysis step wherein the source of the nucleic acid is determined.

The limitations of such pre-existing array technologies are evident in WO 98/44151 and WO 00/18957, for example, which both describe methods of nucleic acid amplification which allow amplification products to be immobilised on a solid support in order to form arrays comprised of clusters or "colonies" formed from a plurality of identical immobilised polynucleotide strands and a plurality of identical immobilised complementary strands. The nucleic acid molecules present in DNA colonies on the clustered arrays prepared according to these methods can provide templates for sequencing reactions, but to date only a single sequencing read can be obtained from each immobilised strand in each colony. The methods described herein allow for two or more reads on a single amplified strand, thereby allowing analysis of the tag sequence independently of the target sequence, and also allowing the possibility of a paired read from the opposite end of each fragment as well as a third read for the tag sequence. The methods also allow the preparation of constructs where the tag sequence is adjacent to the target sequence, and the tag and target are read in a single read, which again can be turned into a paired read methodology if desired.

As will be apparent to the skilled reader, references herein to a particular nucleic acid sequence may, depending on the context, also refer to nucleic acid molecules which comprise the nucleic acid sequence. Sequencing of a target fragment means that a read of the chronological order of bases is established. The bases do not need to be contiguous, although this is preferred, nor does every base on the entire fragment have to be sequenced.

The following passages describe different aspects of the invention in greater detail. Each aspect of the invention may be combined with any other aspect or aspects of the invention unless clearly indicated to the contrary. In particular, any feature indicated as being preferred or advantageous may be combined with any other feature of features indicated as being preferred or advantageous.

The terms 'target nucleic acid sequence', 'target nucleic acid molecule', 'target nucleic acid' and 'target nucleic acid fragment' may be used interchangeably to refer to nucleic acid molecules that it is desired to sequence on an array according to the invention. The target nucleic acid may be essentially any nucleic acid of known or unknown sequence. It may be, for example, a fragment of genomic DNA or cDNA. Sequencing may result in determination of the sequence of the whole, or a part of the target molecule. The targets can be derived from a primary nucleic acid sample that has been randomly fragmented. The targets can be processed into templates suitable for amplification by the placement of universal amplification sequences at the ends of each target fragment. The targets can also be obtained from a primary RNA sample by reverse transcription into cDNA.

As used herein, the term 'polynucleotide' refers to deoxyribonucleic acid (DNA), but where appropriate the skilled artisan will recognise that the method may also be applied to ribonucleic acid (RNA). The terms should be understood to include, as equivalents, analogs of either DNA or RNA made from nucleotide analogs and to be applicable to single stranded (such as sense or antisense) and double stranded polynucleotides. The term as used herein also encompasses cDNA, that is complementary or copy DNA produced from an RNA template, for example by the action of reverse transcriptase.

The primary polynucleotide molecules may originate in double-stranded DNA (dsDNA) form (e.g. genomic DNA fragments, PCR and amplification products and the like) or may have originated in single-stranded form, as DNA or RNA, and been converted to dsDNA form. By way of example, mRNA molecules may be copied into double-stranded cDNAs suitable for use in the method of the invention using standard techniques well known in the art. The precise sequence of the primary polynucleotide molecules is generally not material to the invention, and may be known or unknown.

In a particular embodiment, the primary polynucleotide molecules are RNA molecules. In an aspect of this embodiment, RNA isolated from specific samples is first converted to double-stranded DNA using techniques known in the art. In accordance with the present method, the double-stranded DNA is then tagged or indexed with a sample specific tag. Different preparations of such double-stranded DNA comprising sample specific tags can be generated, in parallel, from RNA isolated from different specific samples. Subsequently, different preparations of double-stranded DNA comprising different sample specific

tags can be mixed, sequenced *en masse*, and the identity of each sequenced fragment determined with respect to the sample from which it was isolated/derived by virtue of the presence of a sample specific tag.

In a particular embodiment, the primary polynucleotide molecules are DNA molecules. More particularly, the primary polynucleotide molecules represent the entire genetic complement of an organism, and are genomic DNA molecules which include both intron and exon sequences (coding sequence), as well as non-coding regulatory sequences such as promoter and enhancer sequences. Although it could be envisaged that particular sub-sets of polynucleotide sequences or genomic DNA could also be used, such as, for example, particular chromosomes. Yet more particularly, the sequence of the primary polynucleotide molecules is not known. Still yet more particularly, the primary polynucleotide molecules are human genomic DNA molecules. The DNA target molecules may be treated chemically or enzymatically either prior to, or subsequent to any random fragmentation processes, and prior to or subsequent to the ligation of the adaptor sequences.

Methods of the invention are especially useful for the parallel sequencing of a large number of relative small samples in the same experiment. In a sequencing platform with a high degree of parallelism, it may be possible to analyse 50-100 million reads in the same experiment, each to a length of 50 bases or more. Thus it may easily be possible to determine more than 5 billion base pairs of sequence from a single experiment. For a genome of 100,000-1 million base pairs, this represents a much larger coverage of each base pair than is actually needed for accurate sequencing. This feature is even more pronounced for viral or mitochondrial

genomes or collections of PCR fragments that may be 10,000 base pairs or less. For a 10 kB sample, sequenced in a run of 5 billion base pairs, each base pair will appear on average 500,000 times. For accurate sequencing, it may be desirable for each base pair to appear only 20 times, and thus it will be possible to sequence 25,000 10kB fragments in the sequencing run. The use of 8 base pair tags gives a possibility of 4^8 (65536) different samples in the same experiment. For the sake of clarity, sequencing of the tags can be performed in a separate read from the sequencing of the target fragments, so the read length or accuracy of the read from the sample is not lowered by the need to sequence the tags.

For genome sizes of 1 mB, at a depth of 20X, it is possible to sequence 250 samples in a 5 billion base pair experiment. The 250 samples can be coded by as few as a four base tag, but extra bases may be added to ensure accurate analysis. Thus, the primary polynucleotide molecules may be a genome of any organism, but the methods of the present invention may be used to particular advantage when analysing smaller genomes, such as those of bacteria, viruses, mitochondria and the like, which have genome sizes in the regions of a few million base pairs or lower.

Random fragmentation refers to the fragmentation of a polynucleotide molecule in a non-ordered fashion by enzymatic, chemical or mechanical means. Such fragmentation methods are known in the art and utilise standard methods (Sambrook and Russell, Molecular Cloning, A Laboratory Manual, third edition). For the sake of clarity, generating smaller fragments of a larger piece of nucleic acid via specific PCR amplification of such smaller fragments is not equivalent to fragmenting the larger piece of nucleic acid

because the larger piece of nucleic acid sequence remains in intact (i.e., is not fragmented by the PCR amplification). Moreover, random fragmentation is designed to produce fragments irrespective of the sequence identity or position of nucleotides comprising and/or surrounding the break. More particularly, the random fragmentation is by mechanical means such as nebulisation or sonication to produce fragments of about 50 base pairs in length to about 1500 base pairs in length, still more particularly 50-700 base pairs in length, yet more particularly 50-400 base pairs in length. Most particularly, the method is used to generate smaller fragments of from 50-150 base pairs in length.

Fragmentation of polynucleotide molecules by mechanical means (nebulization, sonication and Hydroshear for example) results in fragments with a heterogeneous mix of blunt and 3'- and 5'-overhanging ends. It is therefore desirable to repair the fragment ends using methods or kits (such as the Lucigen DNA terminator End Repair Kit) known in the art to generate ends that are optimal for insertion, for example, into blunt sites of cloning vectors. In a particular embodiment, the fragment ends of the population of nucleic acids are blunt ended. More particularly, the fragment ends are blunt ended and phosphorylated. The phosphate moiety can be introduced via enzymatic treatment, for example, using polynucleotide kinase.

In a particular embodiment, the target polynucleotide sequences are prepared with single overhanging nucleotides by, for example, activity of certain types of DNA polymerase such as Taq polymerase or Klenow exo minus polymerase which has a nontemplate-dependent terminal transferase activity that adds a single deoxynucleotide, for example, deoxyadenosine (A) to the 3´ ends of, for example, PCR

products. Such enzymes can be utilised to add a single nucleotide 'A' to the blunt ended 3' terminus of each strand of the target polynucleotide duplexes. Thus, an 'A' could be added to the 3' terminus of each end repaired duplex strand of the target polynucleotide duplex by reaction with Taq or Klenow exo minus polymerase, whilst the adaptor polynucleotide construct could be a T-construct with a compatible 'T' overhang present on the 3' terminus of each duplex region of the adaptor construct. This end modification also prevents self-ligation of both vector and target such that there is a bias towards formation of the combined ligated adaptor-target sequences.

The method of the invention utilises nucleic acid sequence tags as markers characteristic of the source of particular target molecules on the array. A nucleic acid sequence tag characteristic of the source is attached to each of the target molecules in each isolated sample before the sample is immobilised for sequencing. The tag is not itself formed by part of the target nucleic acid molecule, but becomes part of the template for amplification. Generally the tag will be a synthetic sequence of nucleotides which is added to the target as part of the template preparation step. Accordingly, a sample specific tag is a nucleic acid sequence tag which is attached to each of the target molecules of a particular sample, the presence of which is indicative of or is used to identify the sample or source from which the target molecules were isolated.

Preferably the nucleic acid sequence tag may be up to 20 nucleotides in length, more preferably 1-10 nucleotides, and most preferably 4-6 nucleotides in length. A four nucleotide tag gives a possibility of multiplexing 256

samples on the same array, a six base tag enables 4096 samples to be processed on the same array.

In the first embodiment of the invention, the tag sequences are introduced in a two step process involving ligation of an adaptor that is common to all samples and amplification with a tagged amplification primer. See Figures 1 and 6. In the second embodiment the tags are on the adaptor, and the amplification can be performed with universal amplification primers. See Figure 2. In the third embodiment, the tags are on the adaptors, but at the terminus of the duplex region such that the tag and target sequences are contiguous. See Figure 3. The third embodiment allows for the sequencing to be performed in a single read, but requires both strands of the adaptor to be modified and therefore requires more oligonucleotides as two unique adaptor strands are needed for each sample.

In each embodiment, the target nucleic acids of each specific fragmented sample are treated by first ligating identical adaptor polynucleotide molecules ('mismatched adaptors', the general features of which are defined below, and further described in copending application US 20070128624, the contents of which are incorporated herein by reference in their entirety) to the 5' and 3' ends of the target polynucleotide duplexes (which may be of known, partially known or unknown sequence) to form adaptor-target constructs and then carrying out an initial primer extension reaction in which extension products complementary to both strands of each individual adaptor-target construct are formed. The resulting primer extension products, and optionally amplified copies thereof, collectively provide a library of template polynucleotides. The term library refers to the collection of target fragments containing known

common sequences at their 3′ and 5′ ends, and may also be referred to as a 3′ and 5′ modified library.

Depending on the embodiment of the invention, the adaptors may be universal for all samples, or one or both strands of the duplexes may carry the tag sequence to code or track the identity of the samples.

The adaptor polynucleotides used in the method of the invention are referred to herein as 'mismatched' adaptors because, as will be explained in detail herein, it is essential that the adaptors include a region of sequence mismatch, i.e., they must not be formed by annealing of fully complementary polynucleotide strands.

Mismatched adaptors for use in the invention are formed by annealing of two partially complementary polynucleotide strands so as to provide, when the two strands are annealed, at least one double-stranded region and at least one unmatched single-stranded region.

The 'double-stranded region' of the adaptor is a short double-stranded region, typically comprising 5 or more consecutive base pairs, formed by annealing of the two partially complementary polynucleotide strands. This term refers to a double-stranded region of nucleic acid in which the two strands are annealed and does not imply any particular structural conformation.

Generally it is advantageous for the double-stranded region to be as short as possible without loss of function. In this context, 'function' refers to the ability of the double-stranded region to form a stable duplex under standard reaction conditions for an enzyme-catalysed nucleic acid ligation reaction, which will be well known to the skilled reader (e.g. incubation at a temperature in the range of 4°C to 25°C in a ligation buffer appropriate for

the enzyme), such that the two strands forming the adaptor remain partially annealed during ligation of the adaptor to a target molecule. It is not absolutely necessary for the double-stranded region to be stable under the conditions typically used in the annealing steps of primer extension or PCR reactions.

Since identical adaptors are ligated to both ends of each target molecule, the target sequence in each adaptor-target construct will be flanked by complementary sequences derived from the double-stranded region of the adaptors. The longer the double-stranded region, and hence the complementary sequences derived therefrom in the adaptor-target constructs, the greater the possibility that the adaptor-target construct is able to fold back and base-pair to itself in these regions of internal self-complementarity under the annealing conditions used in primer extension and/or PCR. It is, therefore, generally preferred for the double-stranded region to be 20 or less, 15 or less, or 10 or less base pairs in length in order to reduce this effect. The stability of the double-stranded region may be increased, and hence its length potentially reduced, by the inclusion of non-natural nucleotides which exhibit stronger base-pairing than standard Watson-Crick base pairs.

It is preferred, but not absolutely essential, for the two strands of the adaptor to be 100% complementary in the double-stranded region. It will be appreciated that one or more nucleotide mismatches may be tolerated within the double-stranded region, provided that the two strands are capable of forming a stable duplex under standard ligation conditions.

Adaptors for use in the invention will generally include a double-stranded region forming the 'ligatable' end

of the adaptor, i.e. the end that is joined to a target polynucleotide in the ligation reaction. The ligatable end of the adaptor may be blunt or, in other embodiments, short 5′ or 3′ overhangs of one or more nucleotides may be present to facilitate/promote ligation. The 5′ terminal nucleotide at the ligatable end of the adaptor is phosphorylated to enable phosphodiester linkage to a 3′ hydroxyl group on the target polynucleotide.

The term 'unmatched region' refers to a region of the adaptor wherein the sequences of the two polynucleotide strands forming the adaptor exhibit a degree of non-complementarity such that the two strands are not capable of fully annealing to each other under standard annealing conditions for a primer extension or PCR reaction. The unmatched region(s) may exhibit some degree of annealing under standard reaction conditions for an enzyme-catalysed ligation reaction, provided that the two strands revert to single stranded form under annealing conditions in an amplification reaction.

It is to be understood that the 'unmatched region' is provided by different portions of the same two polynucleotide strands which form the double-stranded region(s). Mismatches in the adaptor construct can take the form of one strand being longer than the other, such that there is a single stranded region on one of the strands, or a sequence selected such that the two strands do not hybridise, and thus form a single stranded region on both strands. The mismatches may also take the form of 'bubbles', wherein both ends of the adaptor construct(s) are capable of hybridising to each other and forming a duplex, but the central region is not. The portion of the strand(s) forming the unmatched region are not annealed under conditions in

which other portions of the same two strands are annealed to form one or more double-stranded regions. For avoidance of doubt it is to be understood that a single-stranded or single base overhang at the 3' end of a polynucleotide duplex that subsequently undergoes ligation to the target sequences does not constitute an 'unmatched region' in the context of this invention.

The lower limit on the length of the unmatched region will typically be determined by function, for example, the need to provide a suitable sequence for binding of a primer for primer extension, PCR and/or sequencing. Theoretically there is no upper limit on the length of the unmatched region, except that in general it is advantageous to minimise the overall length of the adaptor, for example, in order to facilitate separation of unbound adaptors from adaptor-target constructs following the ligation step. Therefore, it is generally preferred that the unmatched region should be less than 50, or less than 40, or less than 30, or less than 25 consecutive nucleotides in length.

The precise nucleotide sequence of the adaptors is generally not material to the invention and may be selected by the user such that the desired sequence elements are ultimately included in the common sequences of the library of templates derived from the adaptors to, for example, provide binding sites for particular sets of universal amplification primers and/or sequencing primers. Additional sequence elements may be included, for example, to provide binding sites for sequencing primers which will ultimately be used in sequencing of template molecules in the library, or products derived from amplification of the template library, for example on a solid support.

Although the precise nucleotide sequence of the adaptor is generally non-limiting to the invention, the sequences of the individual strands in the unmatched region should be such that neither individual strand exhibits any internal self-complementarity which could lead to self-annealing, formation of hairpin structures, etc. under standard annealing conditions. Self-annealing of a strand in the unmatched region is to be avoided as it may prevent or reduce specific binding of an amplification primer to this strand.

The mismatched adaptors are preferably formed from two strands of DNA, but may include mixtures of natural and non-natural nucleotides (e.g. one or more ribonucleotides) linked by a mixture of phosphodiester and non-phosphodiester backbone linkages. Other non-nucleotide modifications may be included such as, for example, biotin moieties, blocking groups and capture moieties for attachment to a solid surface, as discussed in further detail below.

The adaptor constructs may contain exonuclease resistant modifications such as phosphorothioate linkages. Such modifications reduce the number of adaptor-dimers present in the library, since the two adaptors can not undergo ligation without removal of their non complementary overhangs. The adaptors can be treated with an exonuclease enzyme prior to the ligation reaction with the target, to ensure that the overhanging ends of the strands can not be removed during the ligation process. Treatment of the adaptors in this manner reduces the formation of the adaptor-dimers at the ligation step.

Ligation methods are known in the art and utilise standard methods (Sambrook and Russell, Molecular Cloning, A Laboratory Manual, third edition). Such methods utilise

ligase enzymes such as DNA ligase to effect or catalyse joining of the ends of the two polynucleotide strands of, in this case, the adaptor duplex construct and the target polynucleotide duplexes, such that covalent linkages are formed. The adaptor duplex construct may contain a 5'-phosphate moiety in order to facilitate ligation to the target 3'-OH. The target contains a 5'-phosphate moiety, either residual from the shearing process, or added using an enzymatic treatment step, and has been end repaired, and optionally extended by an overhanging base or bases, to give a 3'-OH suitable for ligation. In this context, joining means covalent linkage of polynucleotide strands which were not previously covalently linked. In a particular aspect of the invention, such joining takes place by formation of a phosphodiester linkage between the two polynucleotide strands, but other means of covalent linkage (e.g. non-phosphodiester backbone linkages) may be used.

Optionally the combined ligated polynucleotide sequences and unligated adaptor polynucleotide constructs may be purified from any components of the ligation reaction, such as enzymes, buffers, salts and the like. Suitable purification methods are known in the art and utilise standard methods (Sambrook and Russell, Molecular Cloning, A Laboratory Manual, third edition).

In a next step according to the invention an amplification reaction is prepared. The contents of an amplification reaction are known by one skilled in the art and include appropriate substrates (such as dNTPs), enzymes (e.g. a DNA polymerase) and buffer components required for an amplification reaction. Generally amplification reactions require at least two amplification primers, often denoted 'forward' and 'reverse' primers (primer oligonucleotides)

that are capable of annealing specifically to a part of the polynucleotide sequence to be amplified under conditions encountered in the primer annealing step of each cycle of an amplification reaction. In certain embodiments the forward and reverse primers may be identical. Thus the primer oligonucleotides must include an 'adaptor-target specific portion', being a sequence of nucleotides capable of annealing to a part of, that is, a primer-binding sequence, in the polynucleotide molecule to be amplified (or the complement thereof if the template is viewed as a single strand) during the annealing step.

Depending on the embodiment of the invention, the amplification primers may be universal for all samples, or one of the forward or reverse primers may carry the tag sequence that codes for the sample source. The amplification primers may hybridise across the tag region of the ligated adaptor, in which case unique primers will be needed for each sample nucleic acid. The amplification reaction may be performed with more than two amplification primers. In order to prevent the amplification of ligated adapter-adapter dimers, the amplification primers can be modified to contain nucleotides that hybridise across the whole of the ligated adapter and into the ligated template (or the dNTP's attached to the 3' end thereof). This first amplification primer can be modified and treated to help prevent exonucleases digestion of the strands, and thus it may be advantageous to have a first amplification primer that is universal and can amplify all samples rather than modifying and treating each of the tagged primers separately. The tagged primer can be introduced as a sample specific third primer in the amplification reaction, but does not need to be specially modified and treated to lower exonuclease

digestion. The nested PCR approach is shown in figure 6. In the case of this embodiment the third amplification primer that carries the tag needs to contain at sequence the same as at least a portion of the first amplification primer such that it can be used to amplify the duplex resulting from extension of the first amplification primer.

In the context of the present invention, the term 'polynucleotide molecule to be amplified' refers to the original or starting adaptor-target-adaptor sequence added to the amplification reaction. The 'adaptor-target specific portion' in the forward and reverse amplification primers refers to a sequence capable of annealing to the original or initial adaptor-target-adaptor present at the start of the amplification reaction and reference to the length of the 'adaptor-target specific portion' relates to the length of the sequence in the primer which anneals to the starting adaptor-target. It will be appreciated that if the primers contain any nucleotide sequence which does not anneal to the starting adaptor-target in the first amplification cycle then this sequence may be copied into the amplification products (assuming the primer does not contain a moiety which prevents read-through of the polymerase). Hence the amplified template strands produced in the first and subsequent cycles of amplification may be longer than the starting adaptor-target strands.

As the mis-matched adapters can be different lengths, the length of adapter sequence added to the 3' and 5' ends of each strand may be different. The amplification primers may also be of different length to each other, and may hybridise to different lengths of the adapter, and therefore the length added to the ends of each strand can be controlled. In the case of nested PCR, the three or more

amplification primers can be designed to be longer than the primer used to amplify the previous amplicon, so the length of the added nucleotides is fully controllable and may be hundreds of base pairs if desired. In the example shown in figure 6b, the first amplification primer adds 13 bases to the ligated adapter, and the third amplification primer adds a further 27 bases such that one end of the amplicon is 40 bases longer than the short arm of adapter-target construct. The short arm of the adapter is 20 bases in length, meaning that the prepared template comprises the genomic region plus 60 added bases at the end. In figure 6b, the second amplification primer is 25 bases longer than the long arm of adapter, which is 32 bases in length plus the additional T that hybridise across the dATP nucleoside added to the genomic sample. Thus the prepared template comprises the genomic fragment, plus the added dATP, plus 57 known bases. Thus in full, one strand of each template duplex comprises from the 5′ end: 60 known bases, T, the genomic fragment, A, 57 known bases-3′ end. This strand is fully complementary to a sequence: 5′-57 known bases, T, genomic fragment, A, 60 known bases-3′ end. The length 57 and 6 are arbitary, and shown for the purpose of clarification, and should not be taken as limiting. The length of the added sequences may be 20-100 bases or more depending on the desired experimental design.

The forward and reverse primers may be of sufficient length to hybridise to the whole of the adaptor sequence and at least one base of the target sequence (or the nucleotide dNTP added as a 3′-overhang on the target strands). The forward and reverse primers may also contain a region that extends beyond the adaptor construct, and therefore the amplification primers may be at least 20-100 bases in

length. The forward and reverse primers may be of significantly different lengths; for example one may be 20-40 bases, whereas the other one may be 40-100 bases in length. The nucleotide sequences of the adaptor-target specific portions of the forward and reverse primers are selected to achieve specific hybridisation to the adaptor-target sequences to be amplified under the conditions of the annealing steps of the amplification reaction, whilst minimising non-specific hybridisation to any other target sequences present.

Skilled readers will appreciate that it is not strictly required for the adaptor-target specific portion to be 100% complementary, a satisfactory level of specific annealing can be achieved with less than perfectly complementary sequences. In particular, one or two mis-matches in the adaptor-target specific portion can usually be tolerated without adversely affecting specificity for the template. Therefore the term 'adaptor-target specific portion' should not be interpreted as requiring 100% complementarity with the adaptor-target. However, the requirement that the primers do not anneal non-specifically to regions of the adaptor-target other than their respective primer-binding sequences must be fulfilled.

Amplification primers are generally single stranded polynucleotide structures. They may also contain a mixture of natural and non-natural bases and also natural and non-natural backbone linkages, provided that any non-natural modifications do not preclude function as a primer - that being defined as the ability to anneal to a template polynucleotide strand during conditions of the amplification reaction and to act as an initiation point for synthesis of

a new polynucleotide strand complementary to the template strand.

Primers may additionally comprise non-nucleotide chemical modifications, for example phosphorothioates to increase exonuclease resistance, again provided such that modifications do not prevent primer function. Modifications may, for example, facilitate attachment of the primer to a solid support, for example a biotin moiety. Certain modifications may themselves improve the function of the molecule as a primer, or may provide some other useful functionality, such as providing a site for cleavage to enable the primer (or an extended polynucleotide strand derived therefrom) to be cleaved from a solid support.

In an embodiment wherein tags are attached to the adaptors, the amplification can be carried out on either the pooled or unpooled samples. In an embodiment wherein universal adaptors are used, tags must be part of the amplification primers, and therefore, each sample must be amplified independently prior to pooling. The pooled nucleic acid samples can then be processed for sequencing

Preparation of immobilised samples for sequencing

The pooled sample can then be immobilised in preparation for sequencing. Sequencing can be performed as an array of single molecules, or can be amplified prior to sequencing. The amplification can be carried out using one or more immobilised primers. The immobilised primer(s) can be a lawn on a planar surface, or on a pool of beads. The pool of beads can be isolated into an emulsion with a single bead in each 'compartment' of the emulsion. At a concentration of only one template per 'compartment', only a single template is amplified on each bead.

The term 'solid-phase amplification' as used herein refers to any nucleic acid amplification reaction carried out on or in association with a solid support such that all or a portion of the amplified products are immobilised on the solid support as they are formed. In particular, the term encompasses solid-phase polymerase chain reaction (solid-phase PCR) and solid phase isothermal amplification which are reactions analogous to standard solution phase amplification, except that one or both of the forward and reverse amplification primers is/are immobilised on the solid support. Solid phase PCR covers systems such as emulsions, wherein one primer is anchored to a bead and the other is in free solution, and colony formation in solid phase gel matrices wherein one primer is anchored to the surface, and one is in free solution.

Although the invention encompasses 'solid-phase' amplification methods in which only one amplification primer is immobilised (the other primer usually being present in free solution), it is preferred for the solid support to be provided with both the forward and the reverse primers immobilised. In practice, there will be a 'plurality' of identical forward primers and/or a 'plurality' of identical reverse primers immobilised on the solid support, since the amplification process requires an excess of primers to sustain amplification. References herein to forward and reverse primers are to be interpreted accordingly as encompassing a 'plurality' of such primers unless the context indicates otherwise.

As will be appreciated by the skilled reader, any given amplification reaction requires at least one type of forward primer and at least one type of reverse primer specific for the template to be amplified. However, in certain

embodiments the forward and reverse primers may comprise template-specific portions of identical sequence, and may have entirely identical nucleotide sequence and structure (including any non-nucleotide modifications). In other words, it is possible to carry out solid-phase amplification using only one type of primer, and such single-primer methods are encompassed within the scope of the invention. Other embodiments may use forward and reverse primers which contain identical template-specific sequences but which differ in some other structural features. For example one type of primer may contain a non-nucleotide modification which is not present in the other.

In all embodiments of the invention, primers for solid-phase amplification are preferably immobilised by single point covalent attachment to the solid support at or near the 5′ end of the primer, leaving the template-specific portion of the primer free to anneal to its cognate template and the 3′ hydroxyl group free for primer extension. Any suitable covalent attachment means known in the art may be used for this purpose. The chosen attachment chemistry will depend on the nature of the solid support, and any derivatisation or functionalisation applied to it. The primer itself may include a moiety, which may be a non-nucleotide chemical modification, to facilitate attachment. In a particular embodiment, the primer may include a sulphur-containing nucleophile, such as phosphorothioate or thiophosphate, at the 5' end. In the case of solid-supported polyacrylamide hydrogels (as described below), this nucleophile will bind to a bromoacetamide group present in the hydrogel. A more particular means of attaching primers and templates to a solid support is via 5' phosphorothioate attachment to a hydrogel comprised of polymerised acrylamide

and *N*-(5-bromoacetamidylpentyl) acrylamide (BRAPA), as described fully in WO05065814.

Certain embodiments of the invention may make use of solid supports comprised of an inert substrate or matrix (e.g. glass slides, polymer beads, etc) which has been "functionalised", for example by application of a layer or coating of an intermediate material comprising reactive groups which permit covalent attachment to biomolecules, such as polynucleotides. Examples of such supports include, but are not limited to, polyacrylamide hydrogels supported on an inert substrate such as glass. In such embodiments, the biomolecules (e.g. polynucleotides) may be directly covalently attached to the intermediate material (e.g. the hydrogel), but the intermediate material may itself be non-covalently attached to the substrate or matrix (e.g. the glass substrate). The term "covalent attachment to a solid support" is to be interpreted accordingly as encompassing this type of arrangement.

The pooled samples may be amplified on beads wherein each bead contains a forward and reverse amplification primer. In a particular embodiment, the library of templates prepared according to the first, second or third aspects of the invention is used to prepare clustered arrays of nucleic acid colonies, analogous to those described in US20050100900, US7115400, WO 00/18957 and WO 98/44151, the contents of which are incorporated herein by reference in their entirety, by solid-phase amplification and more particularly solid phase isothermal amplification. The terms 'cluster' and 'colony' are used interchangeably herein to refer to a discrete site on a solid support comprised of a plurality of identical immobilised nucleic acid strands and a plurality of identical immobilised complementary nucleic

acid strands. The term 'clustered array' refers to an array formed from such clusters or colonies. In this context the term 'array' is not to be understood as requiring an ordered arrangement of clusters.

The term solid phase, or surface, is used to mean either a planar array wherein primers are attached to a flat surface, for example, glass, silica or plastic microscope slides or similar flow cell devices; beads, wherein either one or two primers are attached to the beads and the beads are amplified; or an array of beads on a surface after the beads have been amplified.

Clustered arrays can be prepared using either a process of thermocycling, as described in WO9844151, or a process whereby the temperature is maintained as a constant, and the cycles of extension and denaturing are performed using changes of reagents. Such isothermal amplification methods are described in patent application numbers WO0246456and US20080009420 (Isothermal methods for creating clonal single molecule arrays), which are incorporated herein by reference in their entirety. Due to the lower temperatures required in the isothermal process, this is particularly preferred.

Use in sequencing/methods of sequencing

The invention also encompasses methods of sequencing amplified nucleic acids generated by solid-phase amplification. Thus, the invention provides a method of nucleic acid sequencing comprising amplifying a pool of nucleic acid templates using solid-phase amplification as described above and carrying out a nucleic acid sequencing reaction to determine the sequence of the whole or a part of at least one amplified nucleic acid strand produced in the

solid-phase amplification reaction, as well as the sequence of the tag that codes for the identity of the sample.

Sequencing can be carried out using any suitable sequencing technique, wherein nucleotides are added successively to a free 3' hydroxyl group, resulting in synthesis of a polynucleotide chain in the 5' to 3' direction. The nature of the nucleotide added is preferably determined after each nucleotide addition. Sequencing techniques using sequencing by ligation, wherein not every contiguous base is sequenced, and techniques such as massively parallel signature sequencing (MPSS) where bases are removed from, rather than added to the strands on the surface are also within the scope of the invention, as are techniques using detection of pyrophosphate release (pyrosequencing). Such pyrosequencing based techniques are particularly applicable to sequencing arrays of beads where the beads have been amplified in an emulsion such that a single template from the library molecule is amplified on each bead.

The initiation point for the sequencing reaction may be provided by annealing of a sequencing primer to a product of the solid-phase amplification reaction. In this connection, one or both of the adaptors added during formation of the template library may include a nucleotide sequence which permits annealing of a sequencing primer to amplified products derived by whole genome or solid-phase amplification of the template library.

Depending on the embodiment of the invention, the tag sequence and fragment sequence may be determined in a single read from a single sequencing primer, or in multiple reads from two sequencing primers. In the case of two reads from two sequencing primers, the 'tag read' and the 'fragment

read' can be performed in either order, with a suitable denaturing step to remove the annealed primer after the first sequencing read is completed. Suitable denaturing steps may include formamide, hydroxide or heat as described elsewhere.

The products of solid-phase amplification reactions wherein both forward and reverse amplification primers are covalently immobilised on the solid surface are so-called 'bridged' structures formed by annealing of pairs of immobilised polynucleotide strands and immobilised complementary strands, both strands being attached to the solid support at the 5' end. Arrays comprised of such bridged structures provide inefficient templates for nucleic acid sequencing, since hybridisation of a conventional sequencing primer to one of the immobilised strands is not favoured compared to annealing of this strand to its immobilised complementary strand under standard conditions for hybridisation.

In order to provide more suitable templates for nucleic acid sequencing, it is preferred to remove substantially all or remove or displace at least a portion of one of the immobilised strands in the 'bridged' structure in order to generate a template which is at least partially single-stranded. The portion of the template which is single-stranded will thus be available for hybridisation to a sequencing primer. The process of removing all or a portion of one immobilised strand in a 'bridged' double-stranded nucleic acid structure may be referred to herein as 'linearisation', and is described in further detail in WO07010251, the contents of which are incorporated herein by reference in their entirety.

Bridged template structures may be linearised by cleavage of one or both strands with a restriction endonuclease or by cleavage of one strand with a nicking endonuclease. Other methods of cleavage can be used as an alternative to restriction enzymes or nicking enzymes, including *inter alia* chemical cleavage (e.g. cleavage of a diol linkage with periodate), cleavage of abasic sites by cleavage with endonuclease (for example 'USER', as supplied by NEB, part number M5505S), or by exposure to heat or alkali, cleavage of ribonucleotides incorporated into amplification products otherwise comprised of deoxyribonucleotides, photochemical cleavage or cleavage of a peptide linker.

It will be appreciated that a linearization step may not be essential if the solid-phase amplification reaction is performed with only one primer covalently immobilised and the other in free solution.

Following the cleavage step, regardless of the method used for cleavage, the product of the cleavage reaction may be subjected to denaturing conditions in order to remove the portion(s) of the cleaved strand(s) that are not attached to the solid support. Suitable denaturing conditions, for example sodium hydroxide solution, formamide solution or heat, will be apparent to the skilled reader with reference to standard molecular biology protocols (Sambrook *et al.*, 2001, *Molecular Cloning, A Laboratory Manual*, 3rd Ed, Cold Spring Harbor Laboratory Press, Cold Spring Harbor Laboratory Press, NY; Current Protocols, eds Ausubel *et al.*). Denaturation results in the production of a sequencing template which is partially or substantially single-stranded. A sequencing reaction may then be initiated by hybridisation of a sequencing primer to the single-stranded portion of the template.

Thus, the invention encompasses methods wherein the nucleic acid sequencing reaction comprises hybridising a sequencing primer to a single-stranded region of a linearised amplification product, sequentially incorporating one or more nucleotides into a polynucleotide strand complementary to the region of amplified template strand to be sequenced, identifying the base present in one or more of the incorporated nucleotide(s) and thereby determining the sequence of a region of the template strand.

One preferred sequencing method which can be used in accordance with the invention relies on the use of modified nucleotides having removable 3′ blocks, for example as described in WO04018497 and US7057026, the contents of which are incorporated herein by reference in their entirety. Once the modified nucleotide has been incorporated into the growing polynucleotide chain complementary to the region of the template being sequenced there is no free 3'-OH group available to direct further sequence extension and therefore the polymerase can not add further nucleotides. Once the nature of the base incorporated into the growing chain has been determined, the 3' block may be removed to allow addition of the next successive nucleotide. By ordering the products derived using these modified nucleotides, it is possible to deduce the DNA sequence of the DNA template. Such reactions can be done in a single experiment if each of the modified nucleotides has a different label attached thereto, known to correspond to the particular base, to facilitate discrimination between the bases added during each incorporation step. Alternatively, a separate reaction may be carried out containing each of the modified nucleotides separately.

The modified nucleotides may carry a label to facilitate their detection. A fluorescent label, for example, may be used for detection of modified nucleotides. Each nucleotide type may thus carry a different fluorescent label, for example, as described in US Provisional Application No. 60/801,270 (Novel dyes and the use of their labelled conjugates), published as WO07135368, the contents of which are incorporated herein by reference in their entirety. The detectable label need not, however, be a fluorescent label. Any label can be used which allows the detection of an incorporated nucleotide.

One method for detecting fluorescently labelled nucleotides comprises using laser light of a wavelength specific for the labelled nucleotides, or the use of other suitable sources of illumination. The fluorescence from the label on the nucleotide may be detected by a CCD camera or other suitable detection means. Suitable instrumentation for recording images of clustered arrays is described in US Provisional Application No. 60/788,248 (Systems and devices for sequence by synthesis analysis), published as WO07123744, the contents of which are incorporated herein by reference in their entirety.

The invention is not intended to be limited to use of the sequencing method outlined above, as essentially any sequencing methodology which relies on successive incorporation of nucleotides into a polynucleotide chain can be used. Suitable alternative techniques include, for example, Pyrosequencing™, FISSEQ (fluorescent in situ sequencing), MPSS and sequencing by ligation-based methods, for example as described is US6306597.

The nucleic acid sample may be further analysed to obtain a second read from the opposite end of the fragment.

Methodology for sequencing both ends of a cluster are described in co-pending applications WO07010252 and PCTGB2007/003798, the contents of which are incorporated by reference herein in their entirety. In one example, the series of steps may be performed as follows; generate clusters, linearise, hybridise first sequencing primer and obtain first sequencing read. The first sequencing primer can be removed, a second primer hybridised and the tag sequenced. The nucleic acid strand may then be 'inverted' on the surface by synthesising a complementary copy from the remaining immobilised primers used in cluster amplification. This process of strand resynthesis regenerates the double stranded cluster. The original template strand can be removed, to linearise the resynthesised strand that can then be annealed to a sequencing primer and sequenced in a third sequencing run.

In the cases where strand resynthesis is employed, both strands must be immobilised to the surface in a way that allows subsequent release of a portion of the immobilised strand. This can be achieved through a number of mechanisms as described in WO07010251, the contents of which are incorporated herein by reference in their entirety. For example, one primer can contain a uracil nucleotide, which means that the strand can be cleaved at the uracil base using the enzymes uracil glycosylase (UDG) which removes the nucleoside base, and endonuclease VIII that excises the abasic nucleotide. This enzyme combination is available as USER™ from New England Biolabs (NEB part number M5505). The second primer may comprise an 8-oxoguanine nucleotide, which is then cleavable by the enzyme FPG (NEB part number M0240). This design of primers give complete control of which primer is cleaved at which point in the process, and also where in the cluster the cleavage occurs. The primers may also be chemically modified, for example with a disulfide or diol

modification that allows chemical cleave at specific locations.

Kits

The invention also relates to kits for use in preparing libraries of tagged polynucleotides using the method of the first, second or third aspects of the invention.

Embodiments of the kit comprise at least a supply of a mismatched adaptor as defined herein, plus a supply of at least one amplification primer which is capable of annealing to the mismatched adaptor and priming synthesis of an extension product, which extension product would include any target sequence ligated to the adaptor when the adaptor is in use.

Particular features of the 'mismatch' adaptors for inclusion in the kit are as described elsewhere herein in relation to other aspects of the invention. The structure and properties of amplification primers are well known to those skilled in the art. Suitable primers of appropriate nucleotide sequence for use with the adaptors included in the kit can be readily prepared using standard automated nucleic acid synthesis equipment and reagents in routine use in the art. The kit may include a supply of one single type of primer or separate supplies (or even a mixture) of two different primers, for example a pair of amplification primers suitable for PCR or isothermal amplification of templates modified with the mismatched adaptor in solution phase and/or on a suitable solid support (i.e. solid-phase amplification). The kit may comprise a mismatch double stranded adapter for ligation to a sample of interest, plus

at least two different amplification primers that carry a different tag sequence, where the tag sequence does not hybridise to the adapter. This kit can be used to amplify at least two different samples where each sample is amplified using a single tagged primer, and then pooled after the individual amplification reactions.

In one embodiment the kit may include supplies of different primer-pairs for use in solution phase and solid phase PCR and more particularly isothermal amplification. In this context the 'different' primer-pairs may be of substantially identical nucleotide sequence but differ with respect to some other feature or modification, such as for example surface-capture moieties, etc. In other embodiments the kit may include a supply of primers for use in an initial primer extension reaction and a different primer-pair (or pairs) for solution and/or solid phase amplification.

Adaptors and/or primers may be supplied in the kits ready for use, or more preferably as concentrates requiring dilution before use, or even in a lyophilised or dried form requiring reconstitution prior to use. If required, the kits may further include a supply of a suitable diluent for dilution or reconstitution of the primers. Optionally, the kits may further comprise supplies of reagents, buffers, enzymes, dNTPs, etc., for use in carrying out PCR or isothermal amplification. Suitable (but non-limiting) examples of such reagents are as described in the Materials and Methods sections of the accompanying Examples. Further components which may optionally be supplied in the kit include 'universal' sequencing primers suitable for sequencing templates prepared using the mismatched adaptors and primers.

The invention will be further understood with reference to the following non-limiting experimental examples:

Examples

The following are examples of general techniques which may be applied in carrying out the method of the invention.

Example 1: Preparation and conventional sequencing of tagged libraries:

The following experimental details describe the complete exposition of one embodiment of the invention as described above.

Two libraries were made, one using purified human BAC DNA (140k human chromosome 6 insert cloned into a pTARBAC vector) and the other purified PhiX174 RF1 DNA (NEB, N3021L). Each library was prepared separately. The DNA was first prepared for ligation to forked adaptors by: fragmentation of the DNA by nebulisation, end repair of the DNA ends to render them blunt-ended and phosphorylated, then the addition of a single 'A' nucleotide onto the 3' ends of the DNA fragments. The ligation reaction was performed with the prepared fragmented DNA and adaptors pre-formed by annealing 'Oligo A' and 'Oligo B' (sequences given below). The same adaptors were used for each library. The product of the reaction was isolated/purified from unligated adaptor by gel electrophoresis. Finally, the product of the ligation reaction was subjected to cycles of PCR to selectively amplify ligated product that contained genomic DNA with adaptor at both ends of the fragments. During the PCR a unique tag was added to each library using a unique PCR

primer, so that the BAC library was indexed with the unique sequence tag 'ATC' and the PhiX library was indexed with the unique sequence tag 'CGA'. See Figure 1 for general schematic of protocol.

Materials and Methods

Step 1) Nebulization

Materials:

- 0.5 ug/ul Human BAC DNA (140k human chromosome 6 insert cloned into a pTARBAC vector)
- 1ug/ul PhiX174 RF1 DNA (NEB, N3021L)
- Nebulization Buffer (53.1 ml glycerol, 42.1 ml water, 3.7 ml 1 M Tris HCl pH7.5, 1.1 ml 0.5 M EDTA)
- TE
- Nebulizers (Invitrogen, K7025-05)
- PCR purification kit columns (Qiagen, 28104)

Procedure:

Mixed 5 µl (5 µg) of PhiX174 DNA with 45 µl of TE and 700 µl 1 of nebulization buffer. Mixed 10 µl (5 µg) of BAC DNA with 40 µl of TE and 700 µl of nebulization buffer. Chilled DNA solutions were each fragmented in a nebulizer on ice for 6 minutes under 32 pounds per square inch (psi) of pressure. The recovered volumes were each purified with a Qiagen PCR purification kit column and eluted in 30 µl of EB.

Step 2) End-repair

Materials:

- Nebulized DNA (from Step 1)
- Water

- T4 DNA ligase buffer with 10mM ATP (10x) (NEB, B0202S)
- dNTPs mix (10mM each) (NEB, N0447S)
- T4 DNA Polymerase (3U/ul) (NEB, M0203L)
- *E. coli* DNA Pol I large fragment (Klenow) (5U/ul) (NEB, M0210S)
- T4 polynucleotide kinase (10U/ul) (NEB, M0201L)
- PCR purification kit columns (Qiagen, 28104)

Procedure:

End repair mix was assembled as follows:

Nebulized DNA	30µl
Water	45µl
T4 DNA ligase buffer with 10mM ATP	10µl
dNTPs	4µl
T4 DNA pol	5µl
Klenow DNA pol	1ul
T4 PNK	5ul
	100µl total

The reaction was incubated for 30 minutes at room temperature. The DNA was purified on a Qiagen column, eluting in 30µl EB.

Step 3) A- tailing Reaction

Materials:

- End repaired DNA (from Step 2)
- Water
- NEB buffer 2 (10x) (NEB, B7002S)
- dATP (1mM) (Amersham-Pharmacia, 272050)
- Klenow fragment (3' to 5' exo minus) (5U/ul) (NEB, M0212B)

- Hot block or PCR machine
- MinElute PCR purification kit column (Qiagen, 28004)

Procedure:

The following reaction mix was assembled:

End repaired DNA	30µl
Water	2ul
NEB buffer 2	5µl
dATP	10µl
Klenow fragment (3' to 5' exo minus)	3µl
	50µl total

The reaction was incubated for 30 minutes at 37°C, then the DNA purified on a Qiagen MinElute column, eluting in 10µl EB.

Step 4) Annealed Adaptors

Materials:

- Oligo A: 5'ACACTCTTTCCCTACACGACGCTCTTCCGATCxT (x = phosphorothioate bond) (SEQ ID No. 1)
- Oligo B: 5'Phosphate-GATCGGAAGAGCGGTTCAGCAGGAATGCCGAG (SEQ ID No. 2)
- 50mM Tris/50mM NaCl pH7.0
- PCR machine

Procedure:

The oligos were mixed together to a final concentration of 15uM each, in 10mM Tris/10mM NaCl pH 7.0. The adaptor strands were annealed in a PCR machine programmed as follows: Ramp at 0.5 °C/sec to 97.5 °C; Hold at 97.5 °C for 150 sec; then a step of 97.5 °C for 2 sec with a temperature drop of 0.1 °C/cycle for 775 cycles.

In order to ensure a sequencing primer is attached to both ends of the adapter for the paired end method, the sequence of oligonucleotide B was changed to:

5' Phosphate-GATCGGAAGAGCACACGTCT 3' (SEQ ID No. 3)

Step 5) Ligation

Materials:

- A-tailed genomic DNA (from Step 3)
- Quick ligase buffer (2x) (NEB, B2200S)
- Annealed adaptor (15uM) (from 4.)
- Quick Ligase (1U/ul) (NEB, M2200L)
- PCR purification kit columns (Qiagen, 28104)

Procedure:

Reaction mix was assembled as follows:

A-tailed genomic DNA	10µl
Quick ligase buffer	25µl
Annealed adaptor	10µl
Quick Ligase	5µl
	50µl total

The reaction was incubated for 15 minutes at room temperature, then the DNA purified on a Qiagen column, eluting in 30µl EB.

Step 6) Gel purification

Materials:

- Ligation reaction (from Step 5)
- Agarose (Biorad, 161-3107)
- TAE (50x)

- Distilled water
- Ethidium bromide (Sigma, E1510)
- Loading buffer (4x) (50 mM Tris pH8, 40 mM EDTA, 40% w/v sucrose)
- Low molecular weight ladder (NEB, N3233L)
- Gel trays and tank. Electrophoresis unit
- Dark reader transilluminator (Clare Chemical Research, D195M)
- Gel extraction kit columns (Qiagen, 28704)

Procedure:

The entire sample from the purified ligation reaction was loaded into one lane of a 2% agarose gel containing ethidium bromide and run at 120V for 60 minutes. The gel was then viewed on a 'White-light' box and fragments from 120bp to 170bp excised and purified with a gel extraction column, eluting in 30µl elution buffer (EB).

Step 7) Exonuclease I Treatment of PCR Primers

Materials:

- Exonuclease I (E. coli) (20 U/ul) (NEB, M0293S)
- Exonuclease I Reaction Buffer (10x) (NEB, M0293S)
- Water
- DNA Primers with a phosphorothioate at the n-1 position
- P6 Bio-Rad columns (Bio-Rad, 732-6221)

Procedure:

DNA Primers with a phosphorothioate at the n-1 position (5 x 85 µl of each Primer (approx 25 µM)) were aliquoted into Eppendorf tubes. 10 µl of 10 X Exonuclease I Reaction Buffer and 5 µl of Exonuclease I was added to each tube.

Each Eppendorf tube was placed in a rack and stored in an oven set at 37 °C for 16 hours. After 16 hr, the tubes were placed on a hotblock set at 80 °C for 2 minutes. Subsequently, the solutions from the Eppendorfs were passed through P6 Bio-Rad columns and spun in a centrifuge at 2000 rpm for 2 minutes. An extra 20 µl of H_20 was added to the columns and the columns respun. The filtered solutions were placed into a SpeedVac® and evaporated until each was at 20 µl, and the fractions combined. The pooled fractions were injected into a reverse phase HPLC system, and the main peak was collected. The collected fractions were evaporated to dryness in a SpeedVac®, 50 µl of water was added and the fraction was subjected again to evaporation to dryness. The resulting pellets were dissolved in 50 µl of water, pooled and a UV measurement taken to determine the concentration of the oligonucleotide.

Step 8) PCR

Materials:

- Gel purified DNA (from Step 6)
- Water
- Phusion master mix (2x) (NEB, F-531L)
- Exonuclease treated universal PCR primer (25uM): 5' AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTCxT (SEQ ID No. 4) 3', where x= phosphorothioate bond (from Step 7)
- Exonuclease treated indexing PCR primer 1 (25uM): 5' CAAGCAGAAGACGGCATACGA**GAT**CGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATCxT (SEQ ID No. 5), where x= phosphorothioate bond (from Step 7)
- Exonuclease treated indexing PCR primer 2 (25uM): 5' CAAGCAGAAGACGGCATACGA**TCG**CGGTCTCGGCATTCCTGCTGAACCGCTCTTC

CGATCxT (SEQ ID No. 6) , where x= phosphorothioate bond (from Step 7)

- PCR machine
- PCR purification kit columns (Qiagen, #28104)

Procedure:

The PCR reaction was prepared as follows:

BAC Library	
Gel purified BAC DNA	1µl
Phusion mastermix	25µl
Universal PCR primer	1µl
Indexing PCR primer 1	1µl
Water	22µl
	50µl total

PhiX174 Library	
Gel purified PhiX174 DNA	1µl
Phusion mastermix	25µl
Universal PCR primer	1µl
Indexing PCR primer 2	1µl
Water	22µl
	50µl total

Thermocycling was carried out in a PCR machine under the following conditions:

- 30 secs @ 98°C
- [10 sec@ 98°C, 30 sec @ 65°C, 30 sec @ 72°C] 18 cycles
- 5 minutes @ 72°C
- Hold @ 4°C

PCR products were purified on a Qiagen column, eluting in 30µl EB. The resulting DNA libraries were ready for

mixing and amplification on a surface amplification platform.

In the case of paired end runs, the human and Phi-X sample were split into 6 portions, and each portion was amplified with the two universal primers plus one of the 12 primers shown below.

Universal LPX primer 5′
GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCxT (SEQ ID No. 7) 3′
Lpx primer 5′
AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCxT (SEQ ID No. 8)
Tag1 primer (CGTGAT) 5′
CAAGCAGAAGACGGCATACGAGATCGTGATGTGACTGGAGTTC (SEQ ID No. 9) 3′
Tag2 primer (ACATCG) 5′
CAAGCAGAAGACGGCATACGAGATACATCGGTGACTGGAGTTC (SEQ ID No. 10) 3′
Tag3 primer (GCCTAA) 5′
CAAGCAGAAGACGGCATACGAGATGCCTAAGTGACTGGAGTTC (SEQ ID No. 11) 3′
Tag4 primer (TGGTCA) 5′
CAAGCAGAAGACGGCATACGAGATTGGTCAGTGACTGGAGTTC (SEQ ID No. 12) 3′
Tag5 primer (CACTGT) 5′
CAAGCAGAAGACGGCATACGAGATCACTGTGTGACTGGAGTTC (SEQ ID No. 13) 3′
Tag6 primer (ATTGGC) 5′
CAAGCAGAAGACGGCATACGAGATATTGGCGTGACTGGAGTTC (SEQ ID No. 14) 3′

Tag7 primer (GATCTG) 5′
CAAGCAGAAGACGGCATACGAGATGATCTGGTGACTGGAGTTC (SEQ ID No. 15)
3′
Tag8 primer (TCAAGT) 5′
CAAGCAGAAGACGGCATACGAGATTCAAGTGTGACTGGAGTTC (SEQ ID No. 16)
3′
Tag9 primer (CTGATC) 5′
CAAGCAGAAGACGGCATACGAGATCTGATCGTGACTGGAGTTC (SEQ ID No. 17)
3′
Tag10 primer (AAGCTA) 5′
CAAGCAGAAGACGGCATACGAGATAAGCTAGTGACTGGAGTTC (SEQ ID No. 18)
3′
Tag11 primer (GTAGCC) 5′
CAAGCAGAAGACGGCATACGAGATGTAGCCGTGACTGGAGTTC (SEQ ID No. 19)
3′
Tag12 primer (TACAAG) 5′
CAAGCAGAAGACGGCATACGAGATTACAAGGTGACTGGAGTTC (SEQ ID No. 20)
3′

The two universal primers contain a phosphorothioate group (x) and were exonucleases treated and purified prior to amplification. The amplification conditions were as described above, and the 6 tagged human samples, and 6 tagged Phi-X sample were pooled together for sequencing on clusters.

Validation of Libraries by Conventional Sanger Sequencing

Four (4) µl of the libraries were cloned into a plasmid vector (Zero Blunt TOPO PCR cloning kit, Invitrogen #K2800-20) and plated out on agar, according to the manufacturer's instructions. Colonies were picked, mini-prepped and the

cloned fragments sequenced by conventional Sanger sequencing.

16 clones from BAC Library (ATC tag)

1 (204bp) Insert: E.coli 85bp

AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTC CGATCTACTGATTTCATTGCAGCCAAAGGCAAACTTTGGCTGCATCGTTTTA CAGTCGCCATAAGCCTTTCCTCTGTTAAACCGCCTTCTGAGATCGGAAGAGC GGTTCAGCAGGAATGCCGAGACCGATCTCGTATGCCGTCTTCTGCTTG (SEQ ID No. 21)

2 (214bp) Insert: BAC 95bp

AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTC CGATCTATCAATATTGTGAAAATGACCATACTGCCAAAAAAAAACTACAAATT CAATGCAATTTTCATCAAAATACCATCATCATTCTTCACAATATTGATAGATC GGAAGAGCGGTTCAGCAGGAATGCCGAGACCGATCTCGTATGCCGTCTTCT GCTTG (SEQ ID No. 22)

3 (215bp) Insert: BAC 96bp

AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTC CGATCTCTCACTCCTGGCAGAGGGACGTGTGACTAGCCATGGGCCCCTAG GTCTCCAGTTCCTGGGTAGCTTGTATTTTTGAACATCTCCTGTATATTAGTTA GATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGATCTCGTATGCCGT CTTCTGCTTG (SEQ ID No. 23)

4 (147bp) Insert: BAC 28bp

AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTC CGATCTAGTGTAGTTGAGATCTGCCTTAGCAGCAAGATCGGAAGAGCGGTT CAGCAGGAATGCCGAGACCGATCTCGTATGCCGTCTTCTGCTTG (SEQ ID No. 24)

5 (183bp) Insert: BAC 64bp

AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTC
CGATCTAACACATTTCAAAGTTTGGGGCCCTCCTCCTCCCCAAAAAACAAAC
CACAAAAAACAAACAAAAAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAG
ACCGATCTCGTATGCCGTCTTCTGCTTG (SEQ ID No. 25)

6 (170bp) Insert: BAC 59bp

GGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTG
AATGCCTTTTATAGCATTTAATTTTTCCTAAGTATAATTACCAAATAAAAATTG
TATAAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGATCTCGTATG
CCGTCTTCTGCTTG (SEQ ID No. 26)

7 (180bp) Insert: BAC 61bp

AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTC
CGATCTTGGGCCCGGGAGGAGTTTGCCGGGGAGGAGTGGGTTTGGAATCG
GGGTTAAAGGAAAGAGAAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAG
ACCGATCTCGTATGCCGTCTTCTGCTTG (SEQ ID No. 27)

8 (190bp) Insert: BAC 73bp

TGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCC
GATCTAAGATCTATTTCAAATGGACTGTAGATCTAAGTATAAAAGGTAAGAGA
ATAATTATTCTAGAAAGTAAATGTAAGATCGGAAGAGCGGTTCAGCAGGAAT
GCCGAGACCGATCTCGTATGCCGTCTTCTGCTTG (SEQ ID No. 28)

9 (192bp) Insert: BAC 74bp

AATGATACGGCGACCACCAGATCTACACTCTTTCCCTACACGACGCTCTTCC
GATCTGGGAGGCCAAGGTGGGTGGATCACCTGAGATCAGGAGTTCGAGAC
CAGCTGGCCAACATGATGAAACTCTGTCTAGATCGGAAGAGCGGTTCAGCA
GGAATGCCGAGACCGATCTCGTATGCCGTCTTCTGCTTG (SEQ ID No.
29)

10 (185bp) Insert: BAC 66bp

AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTC
CGATCTTGACCATTGTAACCATTAATGTAGACTGCAATGATATGCACTATTTA
CAACCTTTTTTAAGACTCTAGATCGGAAGAGCGGTTCAGCAGGAATGCCGA
GACCGATCTCGTATGCCGTCTTCTGCTTG (SEQ ID No. 30)

11 (199bp) Insert: BAC 80bp

AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTC
CGATCTCTTTGAAGAGCTGGCAGTAGAAGATAAACAGGCTGGGGAAGAAGA
GAAAGTGCTCAAGGAGAAGGAGCAGCAGCAGCAGC
AGATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGATCTCGTATGCCGT
CTTCTGCTTG (SEQ ID No. 31)

12 (212bp) Insert: BAC 93bp

AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTC
CGATCTAGTATTCAACAAGTCTGTCTTTTCCAAGTGTCTTTAAAGACCAGAAA
TACCTGTTTTTAACACACAGGGTTGCAAAATTCAGAGGAGATTGGCAGATCG
GAAGAGCGGTTCAGCAGGAATGCCGAGACCGATCTCGTATGCCGTCTTCTG
CTTG (SEQ ID No. 32)

13 (247bp) Insert: E.coli 128bp

AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTC
CGATCTCTTGAGATGAGTGATGACGGCGCGCTGGAAGTTGCTCGTCGCGCT
CGCGGTACGCCGCGCATTGCCAACCGTCTGCTGCGTCGAGTGCGTGATTTC
GCCGAAGTGAAGCACGATGGCACCATCTCAAGAGATCGGAAGAGCGGTTCA
GCAGGAATGCCGAGACCGATCTCGTATGCCGTCTTCTGCTTG (SEQ ID No.
33)

14 (202bp) Insert: BAC 83bp

AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTC
CGATCTGGGGTTGGTGGAACCCAGATGCCTCCCAGGATTGGTGGGCCCTG

TGGCACTTGTACCTGCTGTTGCTGTTGCTGCTGCTGCTGAGATCGGAAGAG
CGGTTCAGCAGGAATGCCGAGACCGATCTCGTATGCCGTCTTCTGCTTG(SE
Q ID No. 34)

15 (166bp) Insert: BAC 47bp
AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTC
CGATCTATGATAAGGAGCAGGTTTACAGATCATAAGTGCAAAAGCGGGCGA
GAAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGATCTCGTATGC
CGTCTTCTGCTTG(SEQ ID No. 35)

16 (147bp) Insert: BAC 31bp
GATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCG
ATCTCTGATACTGTTGTAACCACCCAATTGGTTCAAGATCGGAAGAGCGGTT
CAGCAGGAATGCCGAGACCGATCTCGTATGCCGTCTTCTGCTTG(SEQ ID
No. 36)

16 clones from PhiX174 Library (CGA tag)

1 (183bp) Insert: PhiX 67bp
AATGATACGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGA
TCTTAAAGCTACCAGTTATATGGCTGTTGGTTTCTATGTGGCTAAATACGTTA
ACAAAAAGTCAGATATGAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAGA
CCGCGATCGTATGCCGTCTTCTGCTTG(SEQ ID No. 37)

2 (174bp) Insert: PhiX 58bp
GATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCG
ATCTCGACCACTCGCGATTCAATCATGACTTCGTGATAAAAGATTGAGTGTG
AGGTTATAACAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGCGAT
CGTATGCCGTCTTCTGCTTG(SEQ ID No. 38)

3 (179bp) Insert: PhiX 62bp

TGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCC
GATCTTCTATGAAGGATGTTTTCCGTTCTGGTGATTCGTCTAAGAAGTTTAAG
ATTGCTGAGGGTCAAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACC
GCGATCGTATGCCGTCTTCTGCTTG (SEQ ID No. 39)

4 (169bp) Insert: PhiX 50bp
AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTC
CGATCTGGTTGCGCCGCCAAAACGTCGGCTACAGTAACTTTTCCCAGCCTC
AATCTAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGCGATCGTAT
GCCGTCTTCTGCTTG (SEQ ID No. 40)

5 (166bp) Insert: PhiX 50bp
TGATACGGCGACCACCGAGATCTACATCTTTCCCTACACGACGCTCTTCCGA
TCTAACAACTGAACGGACTGGAAACACTGGTCATAATCATGGTGGCGAATAA
GAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGCGATCGTATGCC
GTCTTCTGCTTG (SEQ ID No. 41)

6 (164bp) Insert: PhiX 45bp
AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTC
CGATCTAAATACGTTAACAAAAAGTCAGATATGGACCTTGCTGCTAAAGGTA
GATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGCGATCGTATGCCGT
CTTCTGCTTG (SEQ ID No. 42)

7 (194bp) Insert: PhiX 75bp
AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTC
CGATCTGTCAATATAACCAGTAGTGTTAACAGTCGGGAGAGGAGTGGCATTA
ACACCATCCTTCATGAACTTAATCCACTGAGATCGGAAGAGCGGTTCAGCAG
GAATGCCGAGACCGCGATCGTATGCCGTCTTCTGCTTG (SEQ ID No. 43)

8 (198bp) Insert: PhiX 79bp

AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTC
CGATCTGATAGTTTGACGGTTAATGCTGGTAATGGTGGTTTTCTTCATTGCAT
TCAGATGGATACATCTGTCAACGCCGCTAATCAGATCGGAAGAGCGGTTCA
GCAGGAATGCCGAGACCGCGATCGTATGCCGTCTTCTGCTTG (SEQ ID No. 44)

9 (179bp) Insert: PhiX 60bp
AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTC
CGATCTTGAAACCAACATAAACATTATTGCCCGGCGTACGAGGAAGGACGT
CAATAGTCACACAGTAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACC
GCGATCGTATGCCGTCTTCTGCTTG (SEQ ID No. 45)

10 (198bp) Insert: PhiX 79bp
AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTC
CGATCTCCAGCTTCTTCGGCACCTGTTTTACAGACACCTAAAGCTACATCGT
CAACGTTATATTTTGATAGTTTGACGGTTAATAAGATCGGAAGAGCGGTTCA
GCAGGAATGCCGAGACCGCGATCGTATGCCGTCTTCTGCTTG (SEQ ID No. 46)

11 (151bp) Insert: PhiX 32bp
AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTC
CGATCTCAGTTTTGCCGCAAGCTGGCTGCTGAACGCCCAGATCGGAAGAGC
GGTTCAGCAGGAATGCCGAGACCGCGATCGTATGCCGTCTTCTGCTTG (SEQ ID No. 47)

12 (173bp) Insert: PhiX 54bp
AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTC
CGATCTTTTGCTGATGAACTAAGTCAACCTCAGCACTAACCTTGCGAGTCAT
TTCTTTGAAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGCGATCG
TATGCCGTCTTCTGCTTG (SEQ ID No. 48)

13 (151bp) Insert: PhiX 32bp

AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTC
CGATCTAATAGACGCAACGCGAGCAGTAGACTCCTTCTAGATCGGAAGAGC
GGTTCAGCAGGAATGCCGAGACCGCGATCGTATGCCGTCTTCTGCTTG(SE
Q ID No. 49)

14 (168bp) Insert: PhiX 49bp

AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTC
CGATCTGGAGGCCTCCAGCAATCTTGAACACTCATCCTTAATACCTTTCTTTT
TGAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGCGATCGTATGC
CGTCTTCTGCTTG(SEQ ID No. 50)

15 (181bp) Insert: PhiX 62bp

AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTC
CGATCTACTCCTAAGCAGAAAACCTACCGCGCTTCGCTTGGTCAACCCCTCA
GCGGCAAAAATTAAAAAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAGAC
CGCGATCGTATGCCGTCTTCTGCTTG(SEQ ID No. 51)

16 (182bp) Insert: E.coli 63bp

AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTC
CGATCTGAGCCGACATCGAGGTGCCAAACACCGCCGTCGATATGAACTCTT
GGGCGGTATCAGCCTGTTAGATCGGAAGAGCGGTTCAGCAGGAATGCCGA
GACCGCGATCGTATGCCGTCTTCTGCTTG(SEQ ID No. 52)

These results confirm that the library preparation method produces 2 libraries of 'sequenceable' DNA templates. Each library contained a plurality of genomic inserts, each of which was flanked by the two adaptors (AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT (SEQ ID No. 53) and AGATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGNNNTCGTATGCCGTCTTCTGCTT G(SEQ ID No. 54)), required for cluster formation and SBS

sequencing. The insert DNA from each of the libraries aligned to the correct reference, either BAC or PhiX, with a small amount of E.coli contamination in each library. The correct TAG ([redacted] for PhiX or [redacted] for BAC) was present in each library.

Example of Cluster preparation using the mixed PHiX and BAC libraries

Example 2): Acrylamide coating of glass chips

The solid supports used are typically 8-channel glass chips such as those provided by Silex Microsystems (Sweden). However, the experimental conditions and procedures are readily applicable to other solid supports.

Chips were washed as follows: neat Decon for 30 minutes, milliQ H_2O for 30 minutes, NaOH 1N for 15 minutes, milliQ H_2O for 30 minutes, HCl 0.1N for 15 minutes, milliQ H_2O for 30 minutes.

Polymer solution preparation:

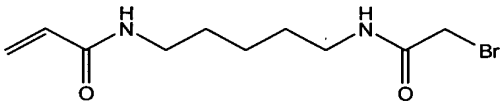
For 10 ml of 2% polymerisation mix.

- 10 ml of 2% solution of acrylamide in milliQ H2O
- 165 µl of a 100mg/ml *N*-(5-bromoacetamidylpentyl) acrylamide (BRAPA) solution in DMF (23.5 mg in 235µl DMF)
- 11.5 µl of TEMED
- 100 µl of a 50 mg/ml solution of potassium persulfate in milliQ H_2O (20mg in 400µl H_2O)

The 10 ml solution of acrylamide was first degassed with argon for 15 minutes. The solutions of BRAPA, TEMED and potassium persulfate were successively added to the

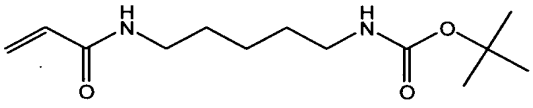
acrylamide solution. The mixture was then quickly vortexed and immediately used. Polymerization was then carried out for 1h 30 minutes at RT. Afterwards the channels were washed with milliQ H_2O for 30 minutes and filled with 0.1 M potassium phosphate buffer for storage until required.

Example 3) Synthesis of *N*-(5-bromoacetamidylpentyl) acrylamide (BRAPA) (1)



(**1**)

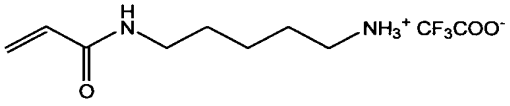
N-Boc-1,5-diaminopentane toluene sulfonic acid was obtained from Novabiochem. The bromoacetyl chloride and acryloyl chloride were obtained from Fluka. All other reagents were Aldrich products.



(**2**)

To a stirred suspension of *N*-Boc-1,5-diaminopentane toluene sulfonic acid (5.2 g, 13.88 mmol) and triethylamine (4.83 ml, 2.5 eq) in THF (120 ml) at 0°C was added acryloyl chloride (1.13 ml, 1 eq) through a pressure equalized dropping funnel over a one hour period. The reaction mixture was then stirred at room temperature and the progress of the reaction checked by TLC (petroleum ether:ethyl acetate 1:1). After two hours, the salts formed during the reaction were filtered off and the filtrate evaporated to dryness. The residue was purified by flash chromatography (neat petroleum ether followed by a gradient of ethyl acetate up to 60%) to

yield 2.56 g (9.98 mmol, 71 %) of product **2** as a beige solid. ^{1}H NMR (400 MHz, d_6-DMSO) : 1.20-1.22 (m, 2H, CH_2), 1.29-1.43 (m, 13H, tBu, 2xCH_2), 2.86 (q, 2H, J = 6.8 Hz and 12.9 Hz, CH_2), 3.07 (q, 2H, J = 6.8 Hz and 12.9 Hz, CH_2), 5.53 (dd, 1H, J = 2.3 Hz and 10.1 Hz, CH), 6.05 (dd, 1H, J = 2.3 Hz and 17.2 Hz, CH), 6.20 (dd, 1H, J = 10.1 Hz and 17.2 Hz, CH), 6.77 (t, 1H, J = 5.3 Hz, NH), 8.04 (bs, 1H, NH). Mass (electrospray+) calculated for $C_{13}H_{24}N_2O_3$ 256, found 279 (256+Na^+).



(**3**)

Product **2** (2.56g, 10 mmol) was dissolved in trifluoroacetic acid:dichloromethane (1:9, 100 ml) and stirred at room temperature. The progress of the reaction was monitored by TLC (dichloromethane:methanol 9:1). On completion, the reaction mixture was evaporated to dryness, the residue co-evaporated three times with toluene and then purified by flash chromatography (neat dichloromethane followed by a gradient of methanol up to 20%). Product **3** was obtained as a white powder (2.43 g, 9 mmol, 90%). ^{1}H NMR (400 MHz, D_2O): 1.29-1.40 (m, 2H, CH_2), 1.52 (quint., 2H, J = 7.1 Hz, CH_2), 1.61 (quint., 2H, J = 7.7 Hz, CH_2), 2.92 (t, 2H, J = 7.6 Hz, CH_2), 3.21 (t, 2H, J = 6.8 Hz, CH_2), 5.68 (dd, 1H, J = 1.5 Hz and 10.1 Hz, CH), 6.10 (dd, 1H, J = 1.5 Hz and 17.2 Hz, CH), 6.20 (dd, 1H, J = 10.1 Hz and 17.2 Hz, CH). Mass (electrospray+) calculated for $C_8H_{16}N_2O$ 156, found 179 (156+Na^+).

To a suspension of product **3** (6.12 g, 22.64 mmol) and triethylamine (6.94 ml, 2.2 eq) in THF (120 ml) was added

bromoacetyl chloride (2.07 ml, 1.1eq), through a pressure equalized dropping funnel, over a one hour period and at -60°C (cardice and isopropanol bath in a Dewar). The reaction mixture was then stirred at room temperature overnight and the completion of the reaction was checked by TLC (dichloromethane:methanol 9:1) the following day. The salts formed during the reaction were filtered off and the reaction mixture evaporated to dryness. The residue was purified by chromatography (neat dichloromethane followed by a gradient of methanol up to 5%). 3.2 g (11.55 mmol, 51 %) of the product **1** (BRAPA) were obtained as a white powder. A further recrystallization performed in petroleum ether:ethyl acetate gave 3g of the product **1**. ^{1}H NMR (400 MHz, d_6-DMSO) : 1.21-1.30 (m, 2H, CH_2), 1.34-1.48 (m, 4H, 2xCH_2), 3.02-3.12 (m, 4H, 2xCH_2), 3.81 (s, 2H, CH_2), 5.56 (d, 1H, J = 9.85 Hz, CH), 6.07 (d, 1H, J = 16.9 Hz, CH), 6.20 (dd, 1H, J = 10.1 Hz and 16.9 Hz, CH), 8.07 (bs, 1H, NH), 8.27 (bs, 1H, NH). Mass (electrospray+) calculated for $C_{10}H_{17}BrN_2O_2$ 276 or 278, found 279 (278+H^+), 299 (276+Na^+).

Example 4) Grafting primers onto surface of SFA coated chip

An SFA coated chip is placed onto a modified MJ-Research thermocycler and attached to a peristaltic pump. Grafting mix consisting of 0.5 µM of a forward primer and 0.5 µM of a reverse primer in 10 mM phosphate buffer (pH 7.0) is pumped into the channels of the chip at a flow rate of 60 µl/minutes for 75 s at 20 °C. The thermocycler is then heated to 51.6 °C, and the chip is incubated at this temperature for 1 hour. During this time, the grafting mix undergoes 18 cycles of pumping: grafting mix is pumped in at 15 µl/minutes for 20 s, then the solution is pumped back and

forth (5 s forward at 15 µl/minutes, then 5 s backward at 15 µl/minutes) for 180 s. After 18 cycles of pumping, the chip is washed by pumping in 5xSSC/5mM EDTA at 15 µl/minutes for 300 s at 51.6 °C. The thermocycler is then cooled to 20 °C.

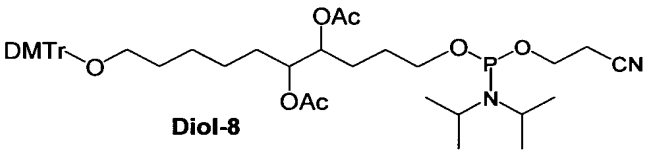
The primers are typically 5'-phosphorothioate oligonucleotides incorporating any specific sequences or modifications required for cleavage. Their sequences and suppliers vary according to the experiment for which they are to be used, and in this case were complementary to the 5'-ends of the template duplex. The DNA sequence used in this process was the pool of the two libraries, which have ends complementary to the grafted primers. The library mix was denatured using sodium hydroxide treatment followed by snap dilution as described.

For some of the experiments detailed, the amplified clusters contained a diol linkage in one of the grafted primers. Diol linkages can be introduced by including a suitable linkage into one of the primers used for solid-phase amplification. Synthesis of the diol phosphoramidite is described in Example 4 below. Products containing such diol linkages can be cleaved using periodate and propanolamine as described, and the resulting single stranded polynucleotides hybridised as described.

The grafted primers contain a sequence of T bases at the 5'-end to act as a spacer group to aid linearisation and hybridization. The sequences of the two primers grafted to the chip are as follows:

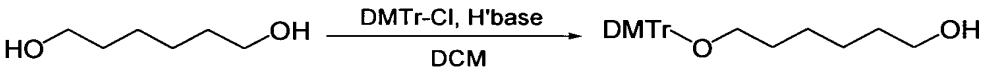
P5diol: 5' PS-TTTTTTTTTT-diol-AATGATACGGCGACCACCGA (SEQ ID No. 55)
P7: 5' PS-TTTTTTTTTTCAAGCAGAAGACGGCATACGA (SEQ ID No. 56)

Example 5: Preparation of diol-phosphoramidite for DNA coupling



Diol-8

Step 1:



Diol-1

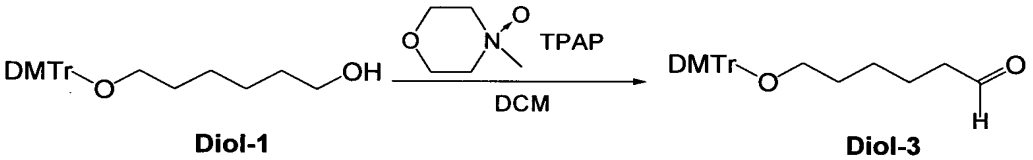
1,6-Hexanediol (Sigma Aldrich 99 %) (14.6 g, 124 mmol), *N,N*-diisopropylethylamine (Hünig's base; Sigma Aldrich; redistilled) (21.6 mL, 124 mmol) is dissolved in anhydrous DCM/ DMF (250/ 50 mL) under N_2. The solution is cooled to 0°C and the first portion of 4, 4'-dimethoxytrityl chloride (DMTr-Cl; Sigma-Aldrich 95 %) (10.5 g, 31 mmol) is added. The reaction mixture is then warmed to room temperature. After stirring for 1 h, the reaction mixture is cooled to 0°C again and the second portion of DMTr-Cl (10.5 g, 31 mmol) is added and then allowed to stir at room temperature for another 2 hours. TLC (EtOAc: petroleum ether 4:6) analysis indicates *ca.* 95% consumption of starting material derivative (DMTr-OH). The reaction is concentrated under reduced pressure and Aq. $NaHCO_3$ (sat.) solution (500 mL) is poured into the residue. The resulting mixture is extracted with petroleum ether/ EtOAc (2:1) (3 x 1000 mL). The

combined organic layers are dried over $MgSO_4$, and concentrated under vacuum. The residue is co-evaporated with xylene (2 x 100 mL) to remove DMF. The reaction mixture, is pre-absorbed on silica gel and subjected to flash chromatography using solvents containing 1 % Et_3N petroleum ether to petroleum ether/ EtOAc (7:3) as eluent. The yield of pale yellow oil is 16.58 g, 64 %, with a further 7.8g (17%) of the bis-tritylated by-product.

TLC: R_f: 0.35 (diol-1); R_f: 0.7 (bis-tritylated by-product) (petroleum ether/ EtOAc 6:4).

^1H NMR (400 MHz, $CDCl_3$): δ 1.32-1.44 (m, 4H, 2x CH_2), 1.54-1.68 (m, 4H, 2x CH_2), 3.06 (t, J = 6.6 Hz, 2H, CH_2O), 3.62-3.68 (m, 2H, CH_2OH), 3.81 (s, 6H, 2x MeO), 6.83-6.85 (m, 4H, Ph), 7.24-7.35 (m, 7H, Ph), 7.45-7.47 (m, 2H, Ph).

Step 2:



Diol-1 Diol-3

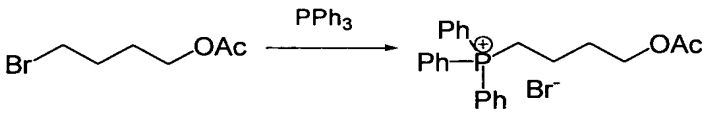
To a solution of Diol-1 (16.6 g, 39.5 mmol) in anhydrous DCM (200 mL), tetrapropylammonium perruthenate (TPAP; Sigma Aldrich 97 %) (277 mg, 0.79 mmol) is added under N_2 atmosphere. The solution is cooled to 0 °C and *N*-methylmopholine *N*-oxide (Sigma Aldrich 97 %) (2.7 g, 23 mmol) is added. The reaction is warmed to room temperature. After 1 hour, the other three portions of *N*-methylmopholine *N*-oxide (3 x 2.0 g, 51.2 mmol) are added within a period of

four hours. TLC (EtOAc: petroleum ether 4:6) indicates the reaction goes to completion. The reaction is quenched with aq. $NaHCO_3$ (sat.) (1000 mL) and extracted to CH_2Cl_2 (4 x 1000 mL). The combined organic layers are dried over $MgSO_4$. The solution is concentrated under reduced pressure. Diol-3, 9.9 g, 60 %, is isolated by flash chromatography using solvents containing 1% Et_3N from petroleum ether to petroleum ether/ EtOAc (6:4) as eluent, as a pale yellow oil.

TLC: R_f: 0.7 (petroleum ether/ EtOAc 6:4).

^1H NMR (400MHz, $CDCl_3$): δ 1.30-1.37 (m, 2H, CH_2), 1.48-1.57 (m, 4H, 2x CH_2), 2.34 (td, J = 1.7 and 7.4 Hz, 2H, CH_2CHO), 2.97 (s, 2H, CH_2O), 3.72 (s, 6H, 2x MeO), 6.73-6.76 (m, 4H, Ph), 7.10-7.26 (m, 7H, Ph), 7.34-7.36 (m, 2H, Ph), 9.67 (t, J = 1.7, 1H, CHO).

Step 3:



Diol-2

A solution of triphenylphosphine (Sigma-Aldrich 99 %, ReagentPlus™). (39.3 g, 150 mmol) and 4-bromobutyl acetate (Sigma-Aldrich)(26 mL, 180 mmol) in anhydrous toluene (300 mL) is heated under reflux for 36 hours under N_2 in an oil-bath (140 °C). During the reflux, oil is precipitated out. The reaction mixture is cooled to room temperature. TLC (petroleum ether/ EtOAc 7:3) analysis of the toluene

solution showed that there is still triphenylphosphine (R_f: 0.8) left. The supernatant is decanted into another round-bottomed flask and concentrated down to the approximate volume of 30 mL. The solution is heated under reflux again for another 12 hours. The supernatant is decanted. The portions of oil are combined together, dissolved in water (500 mL) and extracted with EtOAc (2 x 500 mL). The combined organic layers are back-extracted with water (150 mL). Two lots of aqueous layers are combined, and evaporated under reduced pressure. The resulting residue is co-evaporated with acetonitrile (2 x 100 mL) to give 78.4 g, 95 % yield of a pale yellow oil. NMR indicates that the product is pure, and is used for the next step reaction without further purification.

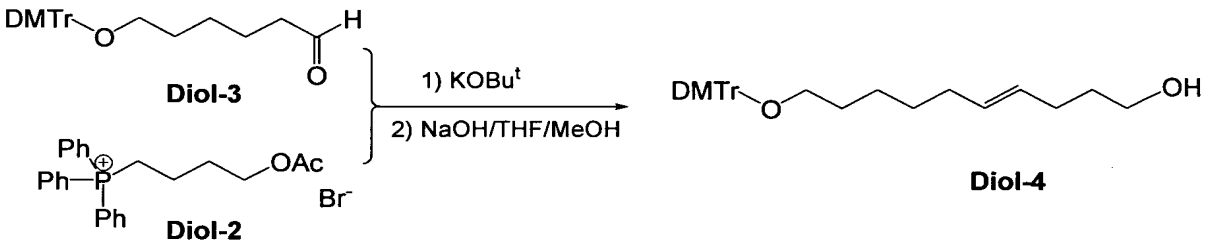
TLC: R_f: 0.0 (petroleum ether/ EtOAc 7:3).

^{1}H NMR (400 MHz, $CDCl_3$): δ 1.63-1.73 (m, 2H, CH_2), 1.94 (s, 3H, 2x CH_3), 2.06-2.16 (m, 2H, CH_2), 3.97-4.05 (m, 2H, CH_2P), 4.11 (t, J = 6.0, 2H, CH_2O), 7.69-7.95 (m, 15H, Ph).

^{31}P NMR (162 MHz, $CDCl_3$): 25.9 ppm.

Mass spec details: LC-MS (Electrospray positive): (M^+) 377.

Step 4:



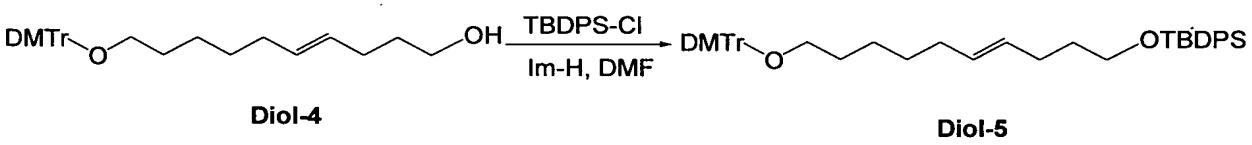
Diol-2 (10.34 g, 22.7 mmol) is weighed into a round-bottomed flask and dissolved with DCM (20 mL). The solution is then evaporated under reduced pressure until it gives a white foam. The flask is then subjected to high vacuum for 24 h. To this flask, anhydrous THF (180 mL) is added under N_2. The resulting suspension is cooled to -78 °C with an acetone-dry ice bath. With vigorous stirring, KOBu^t (3.3 g, 29.5 mmol) is added under N_2. Slowly the colour of the suspension turns orange, and white solids are gradually precipitated out. To this suspension, a solution of diol-3 (dried at 60 °C under high vacuum for 1 h before the reaction),(9.5 g, 22.7 mmol) in THF (50 mL) is added drop wise over half an hour. Acetone-dry ice bath is then removed. The reaction mixture is slowly warmed to room temperature and stirred for another hour. The colour of the reaction mixture turns yellow after the addition of diol-3. The reaction mixture is concentrated down under reduced pressure. The resulting residue is partitioned between DCM (800 mL) and aq. NaCl (sat.) (800 mL). The aqueous layer is extracted with an additional DCM (2 x 800 mL). The organic extractions are combined, dried over $MgSO_4$, filtered, and evaporated under reduced pressure to give yellow oil. The oil is dissolved in THF/MeOH (125/100 mL) and cooled to 0°C. To this solution, NaOH (1M in H_2O, 25 mL) is added. After allowing the reaction to stir for 1 hour, TLC analysis

indicates full consumption of starting material. The reaction mixture is neutralized with acetic acid (1.5 mL). The reaction mixture is concentrated down under reduced pressure. The resulting residue is partitioned between DCM (800 mL) and aq. $NaHCO_3$ (sat.) (800 mL). The aqueous layer is extracted with additional DCM (2 x 800 mL). The organic extractions are combined, dried over $MgSO_4$, filtered, and evaporated to give a pale yellow oil. Diol-4, 6.45 g, 60 % is isolated by flash chromatography using solvents containing 1% Et_3N from petroleum ether to petroleum ether/ EtOAc (5:5) as eluent, as a light yellow oil.

TLC: R_f = 0.45 (petroleum ether/ EtOAc 6:4).

^1H NMR (400 MHz, $CDCl_3$) δ 1.24-1.32 (m, 4H, 2x CH_2), 1.54-1.57 (m, 4H, 2x CH_2), 1.93-1.96 (m, 2H, CH_2), 2.02-2.07 (m, 2H, CH_2), 2.96 (t, J = 6.6 Hz, 2H, CH_2O), 3.54-3.59 (m, 2H, CH_2OH), 3.72 (s, 6H, 2x MeO), 5.29-5.32 (m, 2H, 2x =CH), 6.73-6.77 (m, 4H, Ph), 7.11-7.27 (m, 7H, Ph), 7.36-7.38 (m, 2H, Ph).

Step 5:



Diol-4 Diol-5

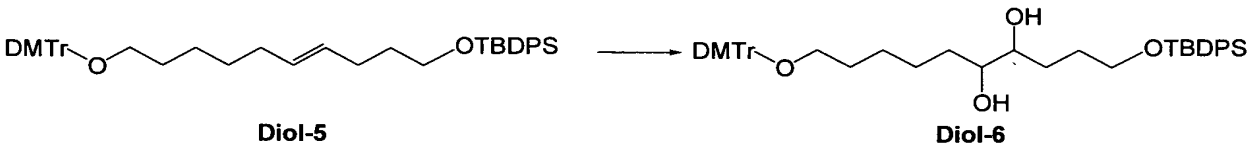
To a solution of Diol-4 (5.68 g, 12 mmol) and imidazole (Sigma Aldrich, 99 %), (1.63 g, 24 mmol) in anhydrous DMF (100 mL), *t*-butyldiphenylsilyl chloride (Sigma Aldrich, 98

%), (4.05 mL, 15.6 mmol) is added drop wise under N_2 atmosphere at room temperature. The reaction is stirred for 1 hour. TLC (petroleum ether/ EtOAc 8:2) indicates that the starting material is fully consumed. A saturated aq. $NaHCO_3$ solution (500 mL) is added to quench the reaction. The resulting mixture is extracted with petroleum ether/EtOAc (2:1) (3 x 500 mL). The organic layers are combined, dried over $MgSO_4$, filtered, and evaporated to give a yellow oil. Diol-5, 8.14 g, 95 % is isolated by flash chromatography using solvents containing 1% Et_3N from petroleum ether to petroleum ether/ EtOAc (9:1) as eluent, as a colourless oil.

TLC: R_f = 0.7 (petroleum ether : EtOAc 8:2).

^1H NMR (400 MHz, $CDCl_3$): δ 0.97 (s, 9H, 3xMe), 1.19-1.30 (m, 4H, 2x CH_2), 1.48-1.55 (m, 4H, 2x CH_2), 1.91-1.95 (m, 2H, CH_2), 2.01-2.06 (m, 2H, CH_2), 2.95 (t, J = 6.6 Hz, 2H, CH_2O), 3.58 (t, J = 6.3 Hz, 2H, CH_2O), 3.70 (s, 6H, 2x MeO), 5.24-5.27 (m, 2H, 2x =CH), 6.72-6.75 (m, 4H, Ph), 7.11-7.37 (m, 15H, Ph), 7.57-7.60 (m, 4H, Ph).

Step 6:



Diol-5 Diol-6

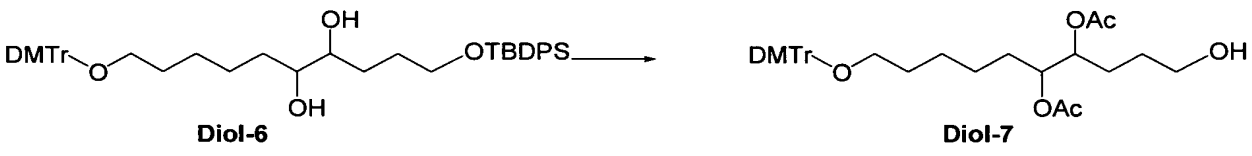
A mixture of diol-5 (9.27 g, 13 mmol), AD-mix-α (Sigma Aldrich), (18.2 g), methanesulfonamide (Sigma Aldrich, 97 %), (1.23 g, 13 mmol), t-BuOH (65 mL) and water (65 mL) is stirred together vigorously at 55 °C for 14 h. The TLC

analysis (petroleum ether : EtOAc 6:4) indicates *ca.* 95 % consumption of the starting material. The reaction mixture is cooled to room temperature, treated with sodium sulfite (15.3 g, 12 mmol), then further stirred for 30 minutes. A saturated aq. $NaHCO_3$ solution (500 mL) is added to the reaction. The resulting mixture is extracted with EtOAc (3 x 500 mL). The organic layers are combined, dried over $MgSO_4$, filtered, and evaporated to give yellow oil. Diol-6, 7.96 g, 82 %, is isolated by flash chromatography (silica gel, Fluka, 70-230 mesh) using solvents containing 1% Et_3N from petroleum ether to petroleum ether/ EtOAc (1:1) as elutant, as a white solid.

TLC: R_f = 0.3 (petroleum ether : EtOAc 6:4).

^1H NMR (400 MHz, $CDCl_3$) δ 1.07 (s, 9 H, 3xMe), 1.41-1.7 (m, 12 H, 6x CH_2), 1.94 (d, J = 4.3 Hz, 1H, OH), 2.94-2.95 (m, 1H, OH), 3.06 (t, J = 6.6 Hz, 2H, CH_2O), 3.61-3.63 (m, 2 H, 2x C*H*OH), 3.73 (t, J = 5.6 Hz, 2H, CH_2O), 3.81 (s, 6H, 2x MeO), 5.24-5.27 (m, 2H, 2x =CH), 6.82-6.85 (m, 4H, Ph), 7.21-7.47 (m, 15 H, Ph), 7.57-7.60 (m, 4 H, Ph).

Step 7:



Diol-6 Diol-7

To a solution of diol-6 (7.96 g, 13 mmol) and DMAP (Sigma-Aldrich ReagentPlus™, 99%). (260 mg, 2.13 mmol) in a

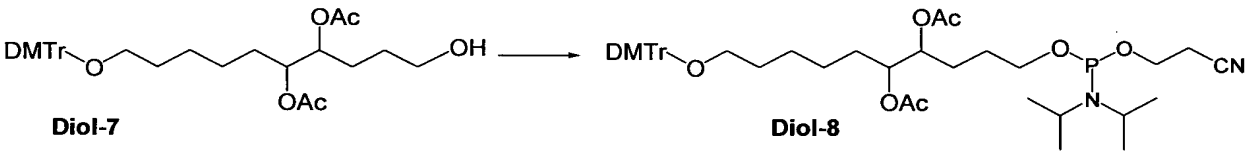
mixture of pyridine (15 mL) and DCM (30 mL), acetic anhydride (Fluka 99%), (2.5 mL, 26.68 mmol) is added at room temperature. TLC analysis (petroleum ether:EtOAc 6:4) indicates full consumption of the starting material after 1 h. The reaction is quenched by saturated aq. $NaHCO_3$ solution (500 mL). After 5 minutes, the mixture is extracted with DCM (3 x 500 mL). The organic layers are combined, dried over $MgSO_4$, filtered, and evaporated. The residue is co-evaporated with toluene (2 x 100 mL). The resulting yellow oil is subjected to a plug of silica gel (50 g, Fluka, 70-230 mesh) to remove DMAP using eluents containing 0.1% Et_3N from petroleum ether to petroleum ether/ EtOAc (7:3) (250 mL each). The combined fractions of product are concentrated to dryness. The resulting colourless oil is dissolved in THF (100 mL) and treated with TBAF (Sigma-Aldrich; 5% wt water), (1 M in THF, 15 mL) at 0 °C. The reaction solution is slowly warmed to room temperature and stirred for 2 hours. TLC analysis (petroleum ether:EtOAc 6:4) indicates that desilylation is completed. The volatile solvent (THF) is evaporated under reduced pressure at low temperature. A saturated aq. $NaHCO_3$ solution (500 mL) is added to the residue. The resulting mixture is extracted with EtOAc (3 x 500 mL). The organic layers are combined, dried over $MgSO_4$, filtered, and evaporated to give yellow oil. Diol-7, 4.2 g, 66 %, is isolated by flash chromatography using solvents containing 1% Et_3N from petroleum ether to petroleum ether/ EtOAc (1:1) as eluent, as a white solid.

TLC: R_f = 0.45 (petroleum ether : EtOAc 1:1).

^1H NMR (400 MHz, $CDCl_3$) δ 1.29-1.33 (m, 4H, 2x CH_2), 1.47-1.63 (m, 8 H, 4x CH_2), 1.99, 2.01 (2s, 6H, 2 MeC(O)), 3.00

(t, *J* = 6.5 Hz, 2H, CH_2O), 3.60-3.64 (m, 2 H, CH_2O), 3.75 (s, 6H, 2x MeO), 4.92-4.97 (m, 2H, 2x C*H*OAc), 6.76-6.80 (m, 4H, Ph), 7.15-7.29 (m, 7 H, Ph), 7.38-7.40 (m, 2 H, Ph).

Step 8:



To a solution of diol-7 (2.08 g, 3.5 mmol) and diisopropylethylamine (Sigma Aldrich), (1.53 ml, 8.75 mmol) in DCM (17 mL), 2-cyanoethyl *N,N*-diisopropylchlorophosphoramidite (1.0 g, 4.2 mmol) is added drop wise at room temperature under N_2. After stirring for 1 hour, TLC analysis (petroleum ether:EtOAc 4:6) indicates the full consumption of the starting material. The solvent (THF) is concentrated under reduced pressure. The resulting residue is subjected to chromatography directly. Diol-8, 2.5 g, 90 %, is isolated by flash chromatography using solvents containing 1% Et_3N from petroleum ether to petroleum ether/ EtOAc (1:1) as eluent, as a colourless syrup.

TLC: R_f = 0.55 (petroleum ether : EtOAc 4:6).

^{1}H NMR (400 MHz, $CDCl_3$) δ 1.09, 1.10, 1.11, 1.12 (4 s, 12 H, $N(CHMe_2)_2$), 1.26-1.31 (m, 4H, 2x CH_2), 1.45-1.56 (m, 8 H, 4x CH_2), 1.95, 1.969, 1.971, 1.98 (4 s, 6H, 2 MeCO), 2.56 (t, *J* = 6.5 Hz, 2H, CH_2CN), 2.95 (t, *J* = 6.5 Hz, 2H, CH_2O), 3.49-3.55 (m, 4 H, CH_2O), 3.72 (s, 6H, 2x MeO), 4.89-4.92 (m, 2H,

2x C*H*OAc), 6.74-6.76 (m, 4H, Ph), 7.13-7.25 (m, 7 H, Ph), 7.34-7.37 (m, 2 H, Ph).

^{31}P NMR (162 MHz, $CDCl_3$): 148.67, 148.69 ppm.

Example 6) Preparation of clusters by isothermal amplification:

Step 1: Hybridisation and amplification

The DNA sequence used in the amplification process is the mixture of the two libraries prepared in Example 1, which have ends complementary to the grafted primers. The duplex DNA (1 nM) is denatured using 0.1 M sodium hydroxide treatment followed by snap dilution to the desired 0.2-2 pM 'working concentration' in 'hybridization buffer' (5 x SSC / 0.1% Tween).

Surface amplification was carried out by isothermal amplification using an MJ Research thermocycler, coupled with an 8-way peristaltic pump Ismatec IPC ISM931 equipped with Ismatec tubing (orange/yellow, 0.51 mm ID). A schematic of the instrument is shown in figure 10.

The single stranded template (denatured as indicated above) is hybridised to the grafted primers immediately prior to the amplification reaction, which thus begins with an initial primer extension step rather than template denaturation. The hybridization procedure begins with a heating step in a stringent buffer to ensure complete denaturation prior to hybridisation. After the hybridization, which occurs during a 20 minutes slow cooling step, the flowcell was washed for 5 minutes with a wash buffer (0.3 x SSC / 0.1% Tween).

A typical amplification process is presented in Table 1, which details the flow volumes per channel.

Table 1. Template Hybridization and 1st Extension

Step	Description	T (°C)	Time (sec)	Flow rate (µl/min)	Pumped V (µl)
1	Pump Hybridization pre-mix	20	120	60	120
2	Pump Hybridization mix	98.5	300	15	75
3	Remove bubbles	98.5	10	100	16.7
4	Stop flow and hold T	98.5	30	static	0
5	Slow cooling	98.5-40.2	19.5 min	static	0
6	Pump wash buffer	40.2	300	15	75
7	Pump amplification pre-mix	40.2	200	15	50
8	Pump amplification mix	40.2	75	60	75
9	First Extension	74	90	static	0
10	cool to room temperature	20	0	static	0

Table 2. Isothermal Amplification

Step	Description	T (°C)	Time (sec)	Flow rate (µl/min)	Pumped V (µl)
(1) This sequence 35 times	Pump Formamide	60	75	60	75
	Pump Amplification pre-mix	60	75	60	75
	Pump Bst mix	60	95	60	95
	Stop flow and hold T	60	180	static	0
2	Pump wash buffer	60	120	60	120

Details are as follows: Hybridisation pre mix (buffer) = 5 x SSC / 0.1% Tween; Hybridisation mix = 0.1 M hydroxide DNA sample, diluted in hybridisation pre mix; Wash buffer = 0.3 x SSC / 0.1% Tween; Amplification pre mix = 2 M betaine, 20 mM Tris, 10 mM Ammonium Sulfate, 2 mM Magnesium sulfate, 0.1% Triton, 1.3% DMSO, pH 8.8; Amplification mix = 2 M

betaine, 20 mM Tris, 10 mM Ammonium Sulfate, 2 mM Magnesium sulfate, 0.1% Triton, 1.3% DMSO, pH 8.8 plus 200 μM dNTP's and 25 units/mL of Taq polymerase (NEB Product ref M0273L); Bst mix = 2 M betaine, 20 mM Tris, 10 mM Ammonium Sulfate, 2 mM Magnesium sulfate, 0.1% Triton, 1.3% DMSO, pH 8.8 plus 200 μM dNTP's and 80 units/mL of Bst polymerase (NEB Product ref M0275L)

Step 2: Blocking extendable 3'-OH groups

To prepare the blocking pre-mix, 1530 μL of water and 170 μL of 10X blocking buffer (NEB buffer 4; product number B7004S) are mixed for a final volume of 1700 μL. To prepare the blocking mix, 1065.13 μL of blocking pre-mix, 21.12 μL of 125 μM ddNTP mix, and 13.75 μL of TdT terminal transferase (NEB; part no M0252S) are mixed for a final volume of 1100 μL.

To block the nucleic acid within the clusters formed in the flow cell channels, the computer component of the instrumentation flows the appropriate blocking buffer through the flow cell, and controls the temperature as shown in the exemplary embodiments below in Table 3.

Step	Description	T (°C)	Time (sec)	Flow rate (μl/min)	Pumped V (μl)
1	Pump Blocking pre-mix	20	200	15	50
2	Pump Blocking mix	37.7	300	15	75
3	Stop flow and hold T	37.7	20	static	0
4	Cyclic pump Blocking mix and	37.7	8 x (20+180)	15 / static	45

	wait				
5	Pump wash buffer	20	300	15	75

Example 7: Linearisation and hybridization of a sequencing primer

To prepare the linearization mix, 1429 µL of water, 64 mg of sodium periodate, 1500 µL of formamide, 60 µL of 1M Tris pH8, and 6011.4 µL of 3-aminopropanol are mixed for a final volume of 3 mL. The periodate is first mixed with the water while the Tris is mixed with the formamide. The two solutions are then mixed together and the 3-aminopropanol is added to that mixture.

To linearize the nucleic acid within the clusters formed within the flow cell channels, 300 µL per channel of linearisation mix is flowed in at 15 µL/minutes at 20 °C; followed by 75 µL of water at the same flow rate.

To prepare the primer mix, 895.5 µL of hybridization buffer and 4.5 µl of sequencing primer (100 µM) are mixed to a final volume of 900 µL. The sequence of the sequencing primer used in the first reaction was:

5'-ACACTCTTTCCCTACACGACGCTCTTCCGATC (SEQ ID No. 57)

To denature the nucleic acid within the clusters and to hybridize the sequencing primer, the following reagents are flowed through the cell as shown in Table 4:

Step	Description	T (°C)	Time (sec)	Flow rate (µl/min)	Pumped V (µl)
1	Pump 0.1 M NaOH	20	300	15	75
2	Pump TE	20	300	15	75
3	Pump Primer mix	20	300	15	75

4	Hold at 60C	60	900	0	0
5	Pump wash buffer	40.2	300	15	75

Example 8: Sequencing from the target fragment

Sequencing of the clusters from the above illustrative protocol was carried out using modified nucleotides prepared as described in International patent application WO 2004/018493, and labeled with four spectrally distinct fluorophores, as described in US application number 60/801270; filed May 18th 2006. Sequencing of clusters is described in more detail in patent WO06064199. The contents of the above-listed three documents are incorporated herein by reference in their entireties.

A mutant 9°N polymerase enzyme (an exo- variant including the triple mutation L408Y/Y409A/P410V and C223S) was used for the nucleotide incorporation steps.

Incorporation mix, Incorporation buffer (50 mM Tris-HCl pH 8.0, 6 mM MgSO4, 1 mM EDTA, 0.05% (v/v) Tween -20, 50mM NaCl) plus 110nM YAV exo- C223S, and 1μM each of the four labeled modified nucleotides, was applied to the clustered templates, and heated to 45°C.

Templates were maintained at 45°C for 30 minutes, cooled to 20°C and washed with Incorporation buffer, then with 5x SSC/0.05% Tween 20. Templates were then exposed to Imaging buffer (100mM Tris pH7.0, 30mM NaCl, 0.05% Tween 20, 50mM sodium ascorbate, freshly dissolved). Templates were scanned in 4 colors at room temperature. Templates were then exposed to sequencing cycles of Cleavage and Incorporation as follows:

Cleavage: Prime with Cleavage buffer (0.1M Tris pH 7.4, 0.1M NaCl and 0.05% Tween 20), 125 µL/channel; 60 µL/minutes.

Heat to 60 °C.

Treat the clusters with Cleavage mix (100 mM TCEP in Cleavage buffer), 75 µL/channel; 60 µL/minutes.

Wait for a total of 15 minutes in addition to pumping fresh cleavage mix, 25 µL/channel; 60 µL/minutes, every 4 minutes.

Cool to 20 °C.

Wash with Enzymology buffer.

Wash with 5XSSC/0.05% Tween 20.

Prime with Imaging buffer.

Scan in 4 colors at RT.

Incorporation: Prime with Incorporation buffer, 125 µL/channel; 60 µL/minutes, Heat to 60°C

Treat with Incorporation mix, 75 µL/channel; 60 µL/minutes.

Wait for a total of 15 minutes in addition to pumping fresh Incorporation mix, 25 µL/channel; 60 µL/minutes, every 4 minutes.

Cool to 20 °C.

Wash with Incorporation buffer, 75 µL/channel; 60 µL/minutes.

Wash with 5XSSC/0.05% Tween 20, 75 µL/channel; 60 µL/minutes

Prime with imaging buffer, 100 µL/channel; 60 µL/minutes

Scan in 4 colors at RT.

Repeat the process of Incorporation and Cleavage for as many cycles as required.

Incorporated nucleotides were detected using a Total Internal Reflection based fluorescent CCD imaging apparatus described in "Systems and Devices for Sequence by Synthesis Analysis," USSN 60/788,248, filed March 31, 2006.

Example 9) Sequencing the tag

The first sequencing primer can be removed, and a second one applied using the same protocol as described in Example 7. The sequence of the second sequencing primer was as follows:

5'-AGATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCG (SEQ ID No. 58)

The sequencing cycles described in Example 8 can be repeated to sequence the tag. In this case, only three cycles of sequencing were performed to analyse the tag. The data in Figure 4 show that all the clusters sequenced aligned against a known sample (PhiX, BAC or E. coli contamination), and that for the samples that aligned correctly, 98% showed the correct tag. The 0.1% errors derive from mis-alignments of the target read, rather than errors in the tag sequences.

Example 10) 12-plex paired reads from a library of fragments

Step 1) Grafting primers onto surface of SFA coated chip

An SFA coated chip is placed onto a modified MJ-Research thermocycler and attached to a peristaltic pump. Grafting mix consisting of 0.5 μM of a forward primer and 0.5 μM of a reverse primer in 10 mM phosphate buffer (pH 7.0) is pumped into the channels of the chip at a flow rate of 60 μl/minutes for 75 s at 20 °C. The thermocycler is then heated to 51.6 °C, and the chip is incubated at this temperature for 1 hour. During this time, the grafting mix undergoes 18 cycles of pumping: grafting mix is pumped in at 15 μl/minutes for 20 s, then the solution is pumped back and forth (5 s forward at 15 μl/minutes, then 5 s backward at 15 μl/minutes) for 180 s. After 18 cycles of pumping, the chip is washed by pumping in 5xSSC/5mM EDTA at 15 μl/minutes for 300 s at 51.6 °C. The thermocycler is then cooled to 20 °C.

The primers are typically 5'-phosphorothioate oligonucleotides incorporating any specific sequences or modifications required for cleavage. Their sequences and suppliers vary according to the experiment for which they are to be used, and in this case were complementary to the 5'-ends of the template duplex. The DNA sequence used in this process was the pool of the twelve libraries, which have ends complementary to the grafted primers. The library mix was denatured using sodium hydroxide treatment followed by snap dilution as described.

The grafted primers contain a sequence of T bases at the 5'-end to act as a spacer group to aid linearisation and hybridization. The sequences of the three primers grafted to the chip are as follows:

Oligo A: 5'-PS-TTTTTTTTTTAATGATACGGCGACCACCGAGAUCTACAC-3'
where **U** = 2-deoxyuridine; (SEQ ID No. 59)

Oligo B: 5'-PS-TTTTTTTTTTCAAGCAGAAGACGGCATAC**Goxo**A-3', where **Goxo** = 8-oxoguanine). (SEQ ID No. 60)

Step 2) Preparation of clusters by isothermal amplification:

The DNA sequence used in the amplification process is the mixture of the twelve libraries prepared in Example 1, which have ends complementary to the grafted primers. The duplex DNA (1 nM) is denatured using 0.1 M sodium hydroxide treatment followed by snap dilution to the desired 0.2-2 pM 'working concentration' in 'hybridization buffer' (5 x SSC / 0.1% Tween).

Surface amplification was carried out by isothermal amplification using a commercially available Solexa/Illumina cluster station as described in WO08002502. The cluster station is essentially a hotplate and a fluidics system for controlled delivery of reagents to a flow cell.

The single stranded template (denatured as indicated above) is hybridised to the grafted primers immediately prior to the amplification reaction, which thus begins with an initial primer extension step rather than template denaturation. The hybridization procedure begins with a heating step in a stringent buffer to ensure complete denaturation prior to hybridisation. After the hybridization, which occurs during a 20 minute slow cooling step, the flowcell was washed for 5 minutes with a wash buffer (0.3 x SSC / 0.1% Tween).

During template hybridization and first extension, a number of solutions/buffers are typically employed, e.g., a solution comprising the DNA samples, a hybridization buffer (5 x SSC / 0.1% tween), a wash buffer (0.3 x SSC / 0.1%

tween), a 2M sodium hydroxide solution, a cluster buffer (200 mM Tris, 100 mM Ammonium Sulfate, 20 mM Magnesium sulfate, 1% Triton, 1.3% DMSO, pH 8.8); an amplification additive (5 M betaine), DNA polymerase, and 10 mM dNTP mix.

To prepare the hybridization mixes, a 0.2 ml strip sample tube and the hybridization buffer are pre-chilled. Using 1.7 ml Eppendorf tube(s), the DNA template(s) are then diluted to 1nM in buffer EB (Qiagen). 1 μL of 2 M NaOH is added to 19 μL of template, vortexed briefly and incubated for 5 minutes at room temperature to denature the DNA template into single strands. The denatured DNA is diluted to working concentration (0.2-2 pM) in pre-chilled hybridization buffer (e.g. for 1 mL of 1pM Hybridization mix, 1 μL of denatured DNA is diluted into 1 mL of pre-chilled hybridization buffer). The volume required depends on the number of channels used - at least 120 μL of hybridization mix per channel is optionally used. Thus, 1 mL of hybridization mix is enough for 8 channels. The samples are vortexed briefly, spun down and aliquoted into the pre-chilled 0.2 ml strip tubes (with no bubbles in the bottom of the tubes) and used immediately.

To prepare the Amplification pre-mix (of sufficient volume for the first extension and 35 cycles of isothermal amplification), 35 mL of H_20 (milliQ), 7 mL of Cluster buffer (200 mM Tris, 100 mM Ammonium Sulfate, 20 mM Magnesium sulfate, 1% Triton, 1.3% DMSO, pH 8.8), and 28 mL of Amplification additive (5 M betaine solution) are mixed to achieve a final volume of 70 mL.

To prepare the first extension Taq mix, 780 μL of Amplification pre-mix, 16 μL of 10 mM dNTPs, and 4 μL of *Taq*

DNA polymerase are mixed together for a final volume of 800 µl.

A typical amplification process is detailed in the following table (Table 1), detailing the flow volumes per channel, controlled automatically by the computer component of the invention.

Table 1. Template hybridization and first extension.

Step	Description	T (°C)	Time (sec)	Flow rate (µl/min)	Pumped V (µl)
1	Pump Hybridization pre-mix	20	120	60	120
2	Pump Hybridization mix	96	300	15	75
3	Remove bubbles	96	6	100	10
4	Stop flow and hold T	96	30	static	0
5	Slow cooling	96-40	1120	static	0
6	Pump wash buffer	40	300	15	75
7	Pump amplification pre-mix	40	280	15	70
8	Pump amplification mix	40	95	60	95
9	First Extension	74	90	static	0
10	cool to room temperature	20	0	static	0

Isothermal amplification at 60 °C using formamide as denaturant

The copied DNA can be isothermally amplified into clusters at 60 °C using formamide as a denaturant. The isothermal amplification (including both temperature control

and reagent control) is overseen by the computer component. Table 2 gives outlines of exemplary script controls. After the isothermal amplification, and optional washing step occur, the nucleic acid of the clusters is ready to be linearized (*see below*).

Table 2. Isothermal amplification

Step	Description	T (°C)	Time (sec)	Flow rate (µl/min)	Pumped V (µl)
(1) This sequence 35 times	Pump Formamide	60	56	30	28
	Pump Amplification pre-mix	60	56	30	28
	Pump Bst mix	60	72	30	36
2	Pump wash buffer	60	280	30	140
3	Pump Storage Buffer	20	380	15	95

Wash buffer = 0.3 x SSC / 0.1% Tween

Amplification pre mix = 2 M betaine, 20 mM Tris, 10 mM Ammonium Sulfate, 2 mM Magnesium sulfate, 0.1% Triton, 1.3% DMSO, pH 8.8

Bst mix = 2 M betaine, 20 mM Tris, 10 mM Ammonium Sulfate, 2 mM Magnesium sulfate, 0.1% Triton, 1.3% DMSO, pH 8.8 plus 200 µM dNTPs and 80 units/mL of Bst polymerase (NEB Product ref M0275L)

Storage Buffer = 5X SSC.

Step 3) Preparation of clusters for first sequencing read

The preparation for read one was performed on the Illumina cluster station. All volumes used in the protocol were 95 ul per lane unless otherwise stated. Linearisation of A-type

surface immobilised oligonucleotides was achieved by incubation with USER enzyme mix (cocktail of Uracil DNA Glycosylase and Endonuclease VIII, NEB #M5505, 10 U/ml, 10 mM KCl, 20 mM Tris pH 8.8, 10 mM (HN4)2SO4, 2 mM MgSO4, 0.1% Triton X-100, 37 °C, 30 minutes). All exposed 3'-OH termini of DNA, either from the extended template or unextended surface oligonucleotides were blocked by dideoxy chain termination using a cocktail of terminal transferase (0.25 U/μl) and a modified polymerase (SBS polymerase as described below) (0.015 mg/ml, 100 μM ddNTP, 50 mM tris, 50 mM NaCl, 6 mM MgSO4, 1 mM EDTA, 0.05 % Tween 20). Blocking was achieved in a two stage protocol, initial incubation at 37 °C for 30 minutes followed by a ramping to 60 °C and incubating the flowcell for a further 15 minutes). Linearised and blocked clusters were washed with 0.3X SSC and storage buffer prior to denaturation with 0.1N NaOH. Denatured clusters were neutralised with TE buffer (10 mM Tris pH 8.0, 1 mM EDTA) and washed with storage buffer. The read 1 specific sequencing primer

(5'- ACACTCTTTCCCTACACGACGCTCTTCCGATC -3'(SEQ ID No. 61), 0.5 μM in hybridisation buffer) was annealed to the clusters by incubation at 60 °C for 15 minutes, followed by a 0.3X SSC wash at 40 °C (ramp rate 1 °C / sec). The flowcell was finally flushed with storage buffer (at 20 °C). Processed flowcells were transferred to the Illumina Genome Analyser for sequencing read 1.

Step 4) Sequencing from the target fragment

Sequencing of the clusters from the above illustrative protocol was carried out using modified nucleotides prepared as described in International patent application WO 2004/018493, and labeled with four spectrally distinct fluorophores, as described in PCT application number PCT/GB2007/001770, published as WO07135368. Sequencing of clusters is described in more detail in patent WO06064199. The contents of the above-listed documents are incorporated herein by reference in their entireties.

A mutant 9°N polymerase enzyme (an exo- variant including the triple mutation L408Y/Y409A/P410V and C223S) (SBS polymerase) was used for the nucleotide incorporation steps.

All processes were conducted as described in the Illumina Genome Analyser operating manual. The flowcell was mounted to the analyser, primed with sequencing reagents: position #1 = incorporation mix (1 μM dNTP mix, 0.015 μg/ml SBS polymerase, 50 mM Tris pH 9.0, 50 mM NaCl, 6 mM MgSO4, 1 mM EDTA, 0.05 % Tween 20); position # 2 = spare (MilliQ water only); position # 3 = scan mix (100 mM Tris pH 7.0, 50 mM sodium acsorbate); position #4 = High salt wash (5x SSC, 0.05 % Tween 20); position # 5 = incorporation buffer (50 mM Tris pH 9.0, 50 mM NaCl, 1 mM EDTA, 0.05 % Tween 20); position #6 = cleavage mix (100 mM TCEP, 100 mM Tris pH 9.0, 100 mM NaCl, 50 mM sodium ascorbate, 0.05% Tween 20); position #7 = cleavage buffer (100 mM Tris pH 9.0, 100 mM NaCl, 0.05% Tween 20); position #8 = spare (single reads) or connected to PE module outlet (paired read experiments). Flowcells were sequenced using standard sequencing recipes for either 27- or 37-cycle experiments. Data was analysed using the standard analysis pipeline.

Incorporation: Prime with Incorporation buffer, 125 µL/channel; 60 µL/minutes, Heat to 60°C

Treat with Incorporation mix, 75 µL/channel; 60 µL/minutes.

Wait for a total of 15 minutes in addition to pumping fresh Incorporation mix, 25 µL/channel; 60 µL/minutes, every 4 minutes.

Cool to 20 °C.

Wash with Incorporation buffer, 75 µL/channel; 60 µL/minutes.

Wash with 5XSSC/0.05% Tween 20, 75 µL/channel; 60 µL/minutes

Prime with imaging buffer, 100 µL/channel; 60 µL/minutes

Scan in 4 colors at RT.

Cleavage: Prime with Cleavage buffer (0.1M Tris pH 7.4, 0.1M NaCl and 0.05% Tween 20), 125 µL/channel; 60 µL/minutes.

Heat to 60 °C.

Treat the clusters with Cleavage mix (100 mM TCEP in Cleavage buffer), 75 µL/channel; 60 µL/minutes.

Wait for a total of 15 minutes in addition to pumping fresh cleavage mix, 25 µL/channel; 60 µL/minutes, every 4 minutes.

Cool to 20 °C.

Wash with Enzymology buffer.

Wash with 5XSSC/0.05% Tween 20.

Repeat the process of Incorporation and Cleavage for as many cycles as required.

Incorporated nucleotides were detected using the Illumina genome analyzer, a Total Internal Reflection based fluorescent CCD imaging apparatus described in "Systems and Devices for Sequence by Synthesis Analysis," USSN 60/788,248, filed March 31, 2006 and corresponding PCT application PCT/US07/07991, published as WO07123744, the contents of which are incorporated herein by reference in their entirety.

Step 5) Sequencing of the Tag

Following the successful completion sequencing the first read on the Genome Analyser, flowcells remained mounted and were prepared for tag sequencing in situ, using the Illumina Paired End module. Temperature control was achieved by using the Genome Analyser peltier. All flow rates were 60 µl/min and 75 µl per lane unless otherwise stated. Clusters were denatured with 0.1 M NaOH to remove the extended sequencing primer from read 1. Hybridisation of the tag sequencing primer

(5'-GATCGGAAGAGCACACGTCTGAACTCCAGTCAC-3') (SEQ ID No. 62), 0.5 µM in hybridisation buffer) as described for read 1. Sequencing of the tag was performed on the Illumina Genome Analyser.

Step 6) Preparation of the clusters for the paired sample read

Following the successful completion sequencing of two reads on the Genome Analyser, flowcells remained mounted and were prepared for paired read 2 *in situ*, using the Illumina Paired End module. Temperature control was achieved by using the Genome Analyser peltier. All flow rates were 60 µl/min and 75 µl per lane unless otherwise stated. Clusters were denatured with 0.1 M NaOH to remove the extended sequencing primer from the tag sequencing. Clusters were 3'-dephosphorylated using T4 polynucleotide kinase (Invitrogen #18004-010, 200 U/ml, 50 mM imidazole pH 6.4, 12 mM MgCl2, 70 µM ADP, 1 mM 2-mercaptoethanol, 37 °C, 30 minutes), prior to re-synthesis of the A-strand achieved using 15 cycles of 60 °C isothermal amplification (same reagents and conditions as described for cluster creation except conducted at 30 µl/min). Clusters were washed before and after resynthesis with 0.3X SSC (150 µl and 245 µl respectively). Linearisation of the B-strand of the re-synthesised clusters was achieved by the excision of 8-oxoguanine from the B-type oligo using Fpg (formamidopyrimidine DNA glycosylase, NEB #M0240, 80 U/ml, 10 mM Bis Tris propane pH 7.0, 10 mM MgCl2, 1mM dithiothreitol, 37 °C, 30 minutes). Blocking was performed as described for read 1 using the same blocking cocktail. Linearised and blocked clusters were denatured prior to hybridisation of the read 2 specific sequencing primer

(5'-GTGACTGGAGTTCAGACGTGTGCTCTTCCGATC-3') (SEQ ID No. 63), 0.5 µM in hybridisation buffer) as described for read 1.

Read 2 of the processed flowcells was subsequently sequenced on the Illumina Genome Analyser.

Data generated from the 12 samples is shown in figure 8. In each case the tag sequence could unambiguously determine the source of the sample. Each sample from the 12 was represented by a reasonable number of clusters, and for each tag read, the large majority (> 90%) of the reads aligned against the source genome identified by the tag.

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 may be used in various combinations. All publications, patents, patent applications, 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, or other document were individually indicated to be incorporated by reference for all purposes.

Claims:

1. A method for sequencing nucleic acid sequences on an array and identifying subsets of nucleic acid sequences on an array, wherein each subset of nucleic acid sequences is isolated from a different source, said method comprising the steps of:

(a) providing at least two samples of randomly fragmented double stranded nucleic acid targets, wherein each of said randomly fragmented double stranded nucleic acid targets is isolated from a different source;

(b) ligating a universal adaptor to the ends of each target fragment of each sample to generate adaptor-target-adaptors of each sample, wherein each of said adaptor-target-adaptors comprises a target fragment flanked by universal adaptor sequences and said universal adaptor comprises a region of double stranded nucleic acid and at least one region of single stranded nucleic acid;

(c) amplifying adaptor-target-adaptors of each sample with two or more sample specific amplification primers to generate amplified nucleic acids, wherein one of said amplification primers comprises a sample specific tag sequence, and wherein amplified nucleic acids of each sample comprise said sample specific tag sequence and said sample specific tag sequence differentiates amplified nucleic acids originating from different samples;

(d) pooling the amplified nucleic acids of different samples;

(e) immobilising the amplified nucleic acids of different samples on an array to generate an array of immobilised fragments;

(f) sequencing the immobilised fragments on the array to determine a sequence read of each immobilised target fragment and identify the sample specific tag sequence of each immobilised fragment, thereby determining both a nucleic acid sequence of said immobilized fragment and identifying the immobilized fragment as a member of a subset of nucleic acids on the array.

2. A method for sequencing nucleic acid sequences on an array and identifying subsets of nucleic acid sequences on an array, wherein each subset of nucleic acid sequences is isolated from a different source, said method comprising the steps of:

(a) providing at least two samples of randomly fragmented double stranded nucleic acid targets, wherein each of said randomly fragmented double stranded nucleic acid targets is isolated from a different source;

(b) ligating a sample specific tagged adaptor to the ends of each target fragment of each sample to generate adaptor-target-adaptors of each sample, wherein each of said adaptor-target-adaptors comprises a target fragment flanked by tagged adaptor sequences and said tagged adaptor comprises a region of double stranded nucleic acid and at least one region of single stranded nucleic acid and a sample specific tag that differentiates adaptor-target-adaptors originating from different samples;

(c) pooling the adaptor-target-adaptors of different samples to generate a pooled sample of adaptor-target-adaptors;

(d) amplifying the pooled sample of adaptor-target-adaptors with a pair of universal amplification primers to generate an amplified pooled sample;

(e) immobilising the amplified pooled sample on an array to generate an array of immobilised fragments;

(f) sequencing the array to determine a sequence read of each immobilised target fragment and identify the sample specific tag sequence of each immobilised fragment, thereby determining both a nucleic acid sequence of said immobilized fragment and identifying the immobilized fragment as a member of a subset of nucleic acids on the array.

3. A method for sequencing nucleic acid sequences on an array and identifying subsets of nucleic acid sequences on an array, wherein each subset of nucleic acid sequences is isolated from a different source, said method comprising the steps of:

(a) providing at least two samples of randomly fragmented double stranded nucleic acid targets, wherein each of said randomly fragmented double stranded nucleic acid targets is isolated from a different source;

(b) ligating a sample specific tagged adaptor to the ends of each target fragment of each sample to generate adaptor-target-adaptors of each sample, wherein each of said adaptor-target-adaptors comprises a target fragment flanked by tagged adaptor sequences and said tagged adaptor comprises a region of double stranded nucleic acid and at least one region of single stranded nucleic acid and a sample specific tag that differentiates adaptor-target-adaptors originating from different samples, and wherein the sample specific tag sequence is attached directly to the target fragment in each adaptor-target-adaptor with no intervening nucleotides between the sample specific tag sequence and target fragment sequences;

(c) amplifying adaptor-target-adaptors of each sample with a pair of amplification primers complementary to the sample specific tag sequence to generate amplified nucleic acids from each sample;

(d) pooling the amplified nucleic acids from each sample to generate a pool of amplified nucleic acid samples;

(e) immobilising the pool of amplified nucleic acid samples on an array to generate an array of immobilised fragments;

(f) sequencing the array to determine a sequence read of each immobilised target fragment and identify the sample specific tag sequence of each immobilised fragment, thereby determining both a nucleic acid sequence of said immobilized fragment and identifying the immobilized fragment as a member of a subset of nucleic acids on the array.

4. The method according to claim 1 or 2, wherein said sequencing is performed in two separate reads.

5. The method according to claim 3, wherein said sequencing of the target fragment and the sample specific tag sequence is performed in a single sequencing read.

6. The method according to claims 1 to 3, wherein the pooled samples are amplified on a solid surface prior to sequencing.

7. The method according to claim 6, wherein said solid surface is a collection of beads.

8. The method according to claim 7, wherein said beads are isolated as single beads in an emulsion.

9. The method according to claims 1 to 3, wherein the array of immobilized fragments is amplified prior to sequencing to generate an amplified array.

10. The method according to claim 9, wherein said amplified array is a clustered array of amplified single molecules.

11. The method according to claim 10, wherein the clustered array is formed by solid-phase nucleic acid amplification.

12. The method according to claim 11, wherein the clustered array is formed by isothermal solid-phase nucleic acid amplification.

13. The method according to any preceding claim, wherein said sequencing involves cycles of ligation with labelled oligonucleotides.

14. The method according to any preceding claim, wherein said sequencing involves cycles of addition of nucleotides.

15. The method according to claim 14, wherein said nucleotides are labelled.

16. The method according to claims 13 or 15, wherein said labels are fluorophores.

17. The method according to any preceding claim, wherein the sequencing step is carried out on an array of immobilised beads.

18. The method according to claim 1 wherein the nucleic acid target fragments can undergo a sequencing reaction from both ends to obtain a paired end read.

19. The method according to claim 18 wherein both ends of the nucleic acid fragment are sequenced before the sequence of the tag is determined.

20. The method according to claim 18 wherein the read from the second end of the fragment occurs after the tag is determined.

21. The method according to claim one that uses two amplification primers per sample, one of which carries the tag.

22. The method according to claim one that uses three amplification primers per sample, one of which carries the tag.

23. A kit comprising a double stranded nucleic acid adapter and at least two different primer polynucleotides that carry a different tag sequence that does not hybridise to the adapter.

24. A kit according to claim 23 further comprising a nucleotide polymerase and four nucleoside triphosphates.

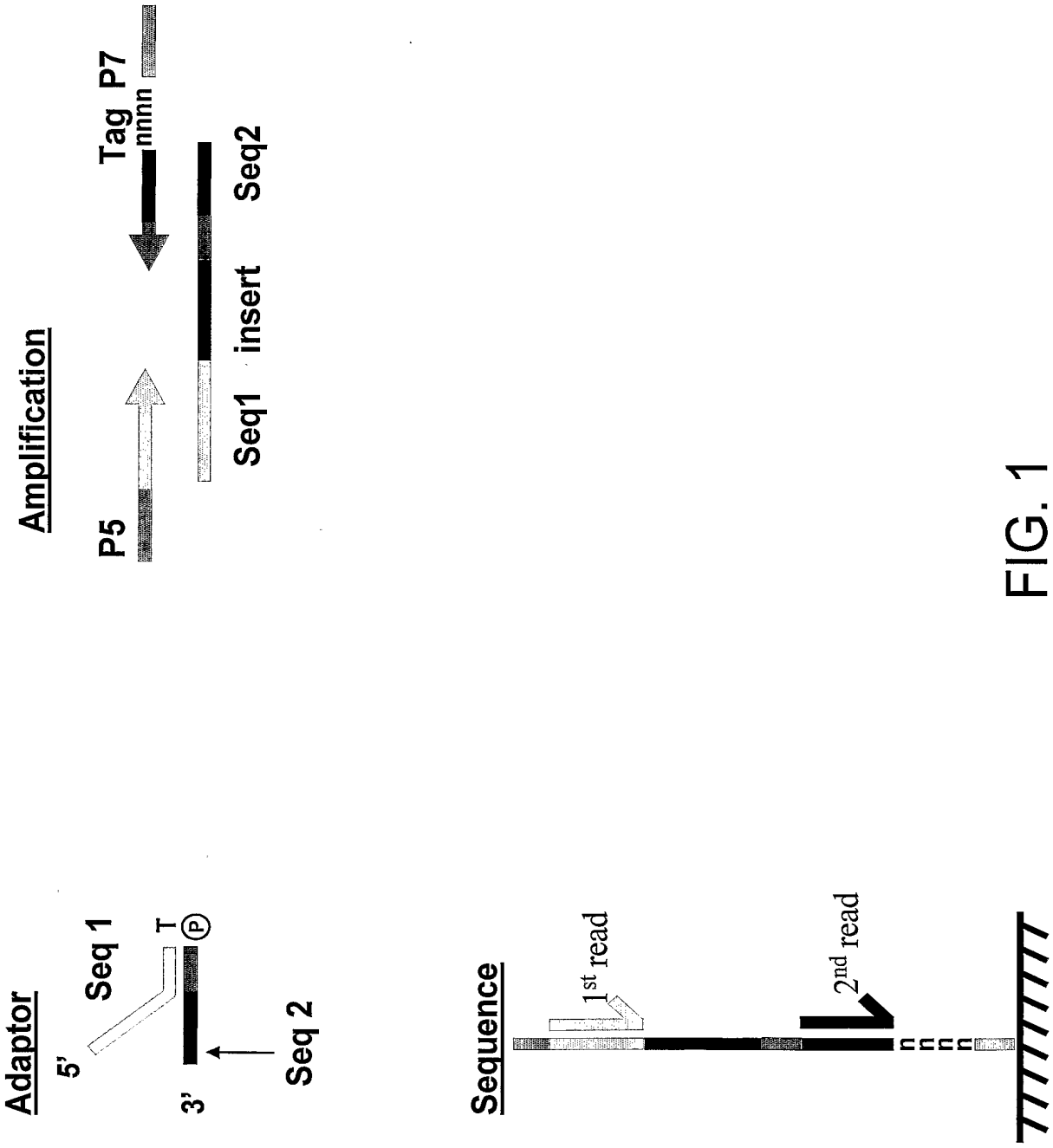


FIG. 1

IDT Exhibit 1018.0293

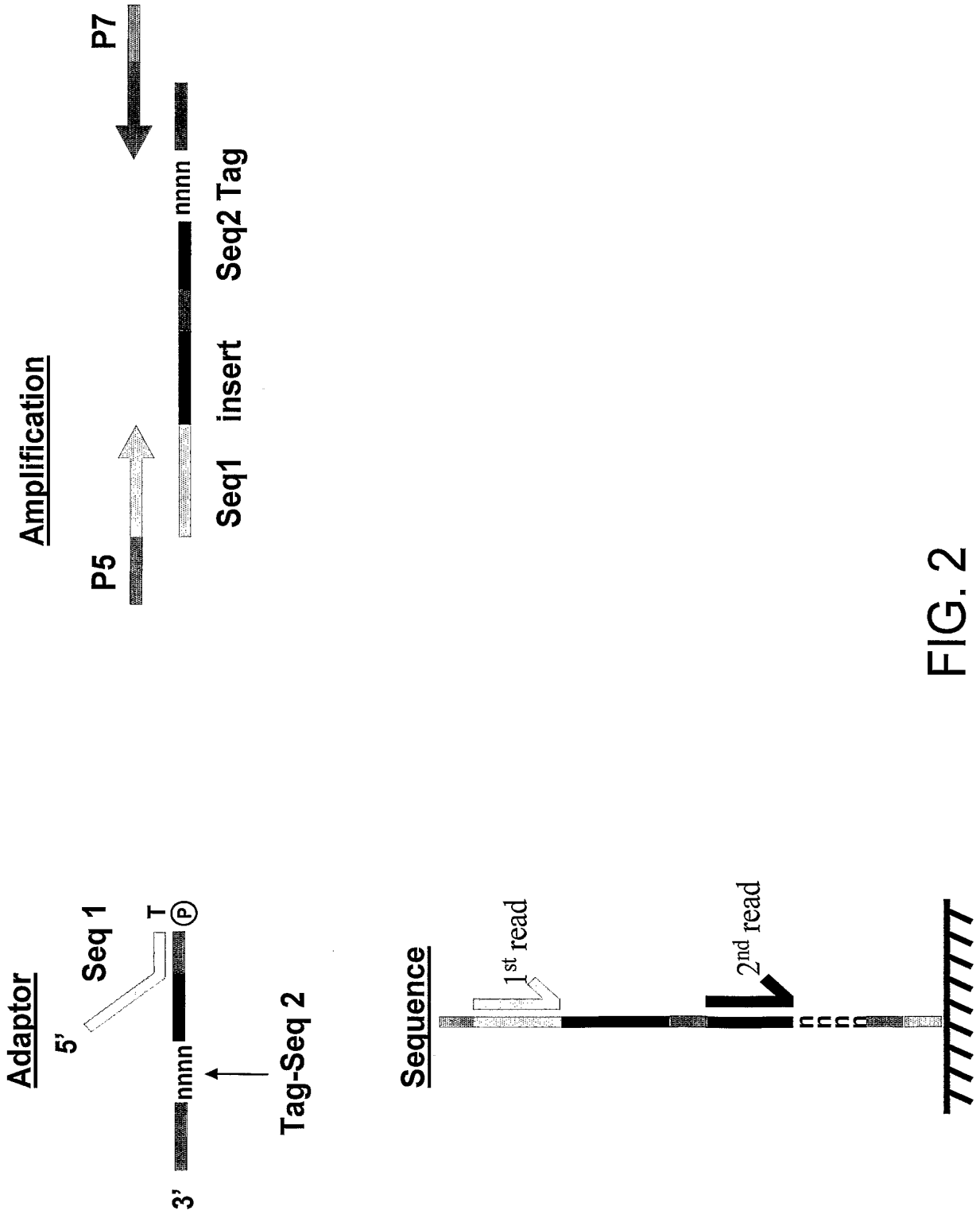


FIG. 2

IDT Exhibit 1018.0294

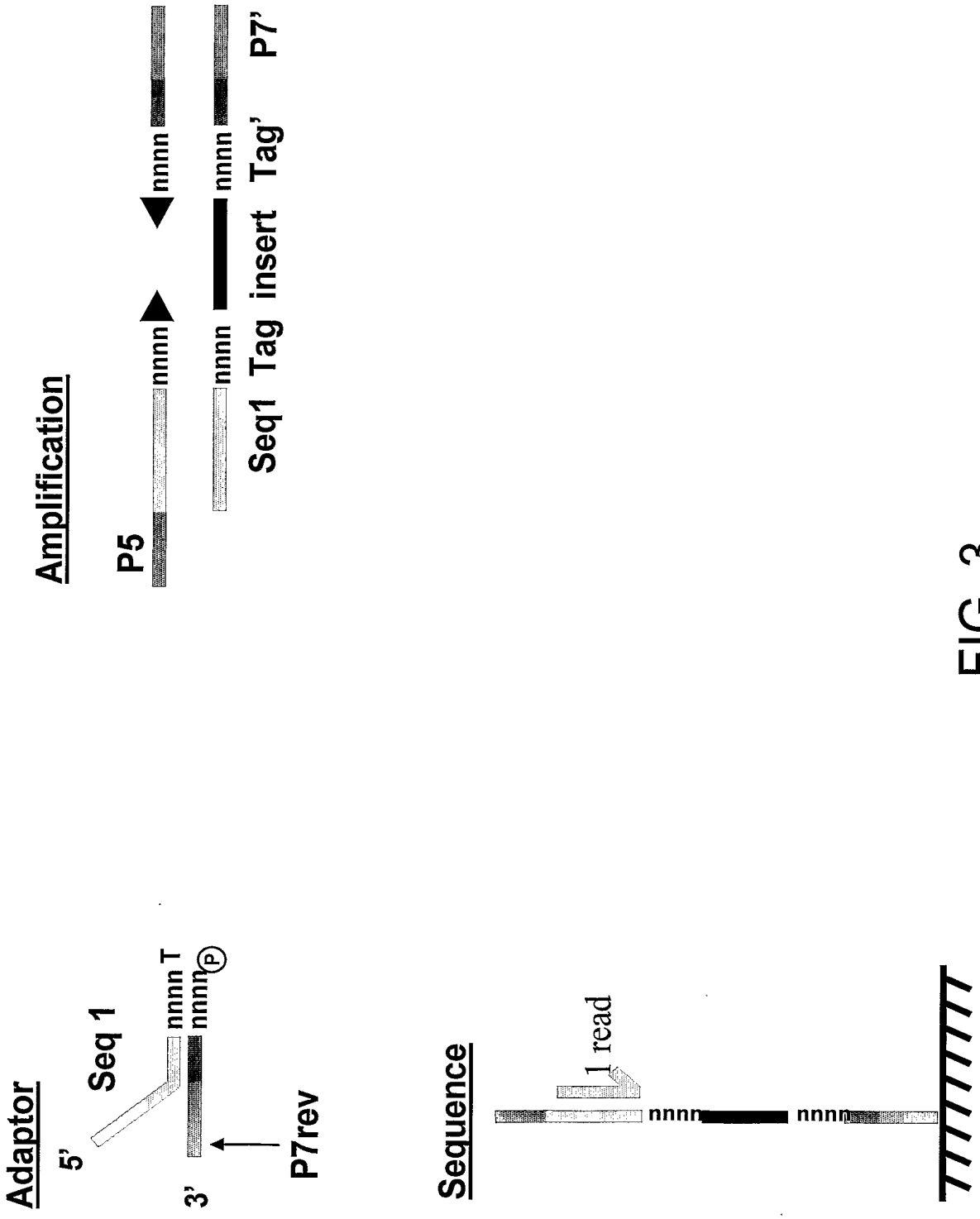

Amplification
P5
nnnn
nnnn
Seq1 Tag insert Tag' P7'
nnnn
Adaptor
5'
Seq 1
nnnn T
nnnn
3'
P7rev
Sequence
1 read
nnnn
nnnn

FIG. 3

- Mixed two libraries and sequenced by SBS in one channel
- Data from a single tile

Reference Sequences	% of Clusters Aligning to Reference
PhiX	68
BAC	29
E.coli	3

	PhiX TAG	BAC TAG	Other TAGs
PhiX	98 %	0.1 %	2 %
BAC	0.1 %	98 %	2 %

- Majority of BAC and PhiX inserts were sequenced with the correct TAG

FIG. 4

IDT Exhibit 1018.0296

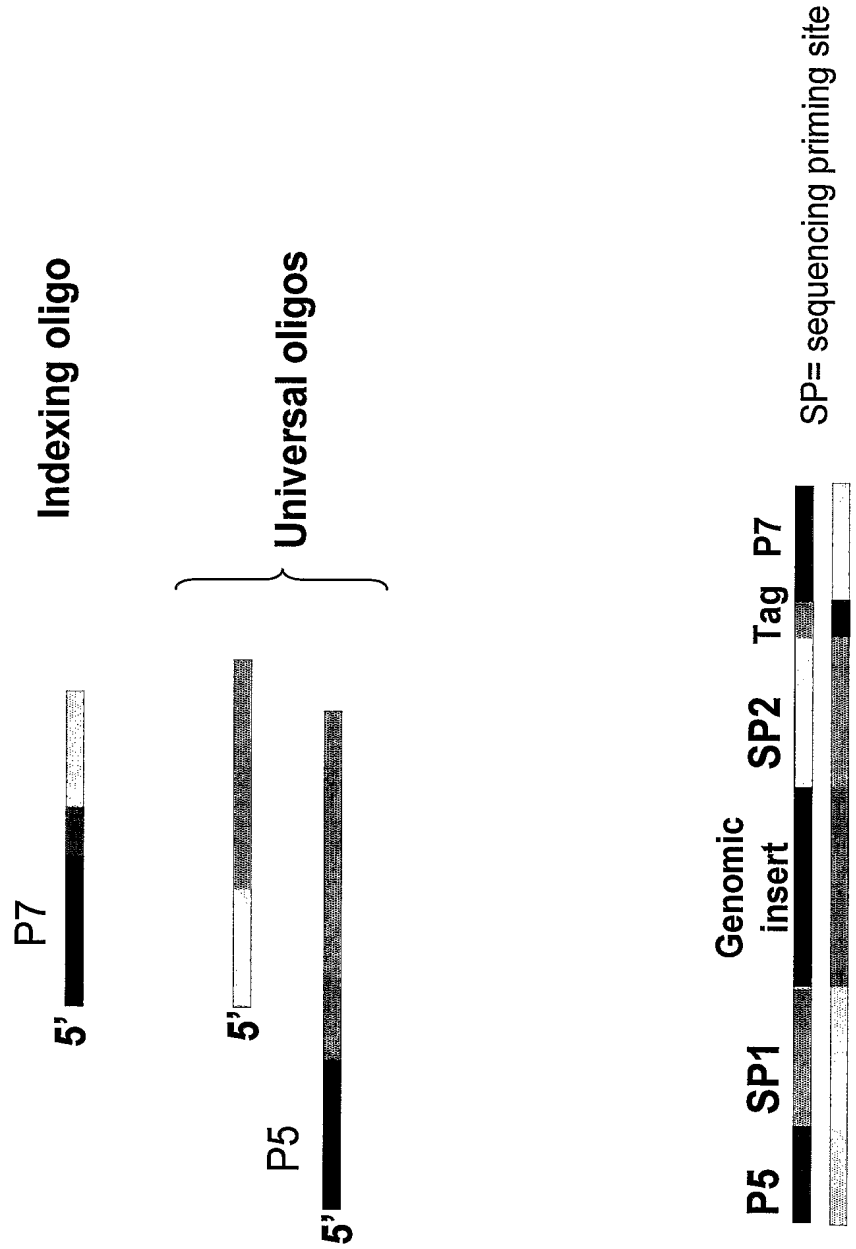


FIG. 5

IDT Exhibit 1018.0297

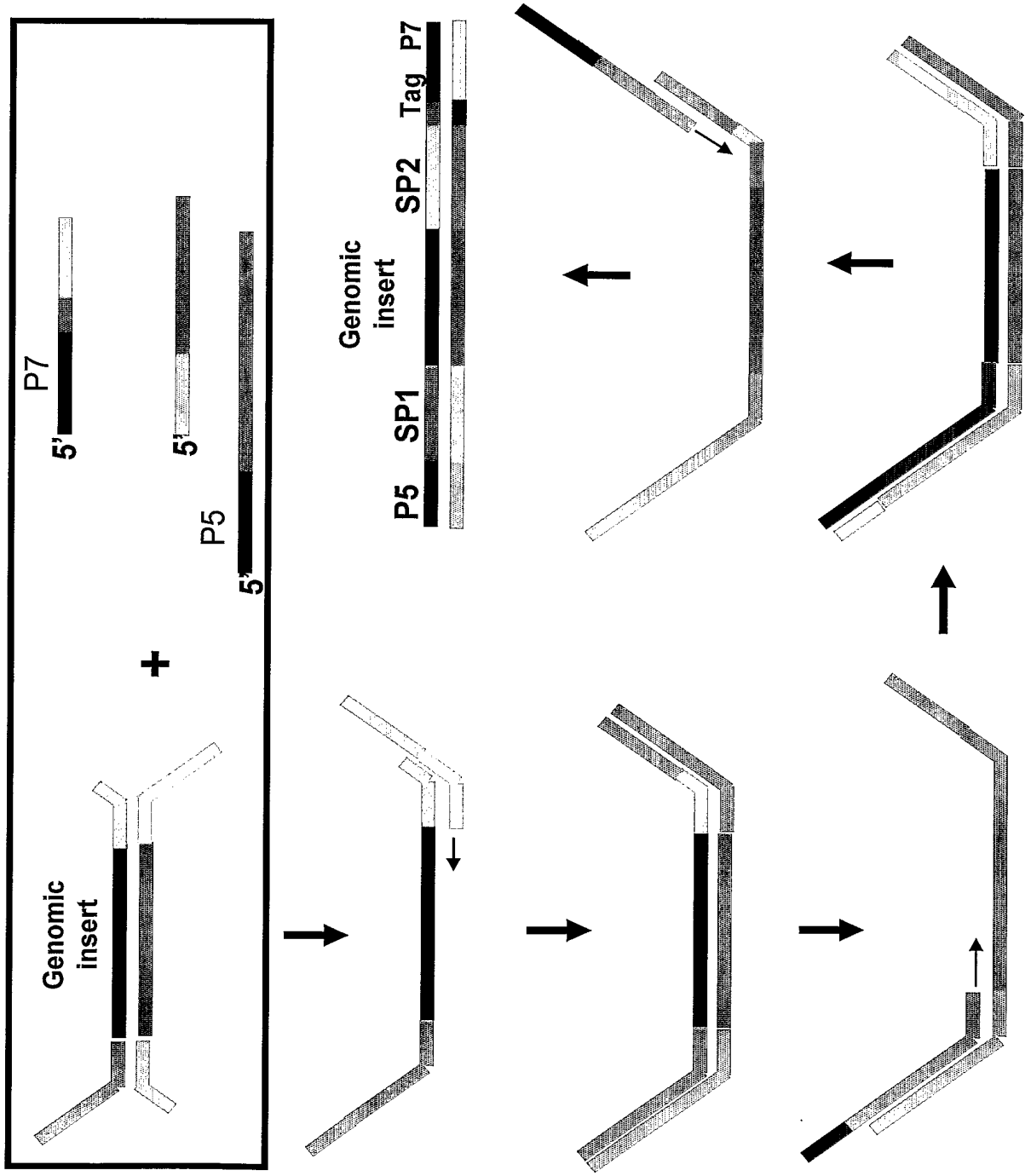


FIG. 6a

IDT Exhibit 1018.0298

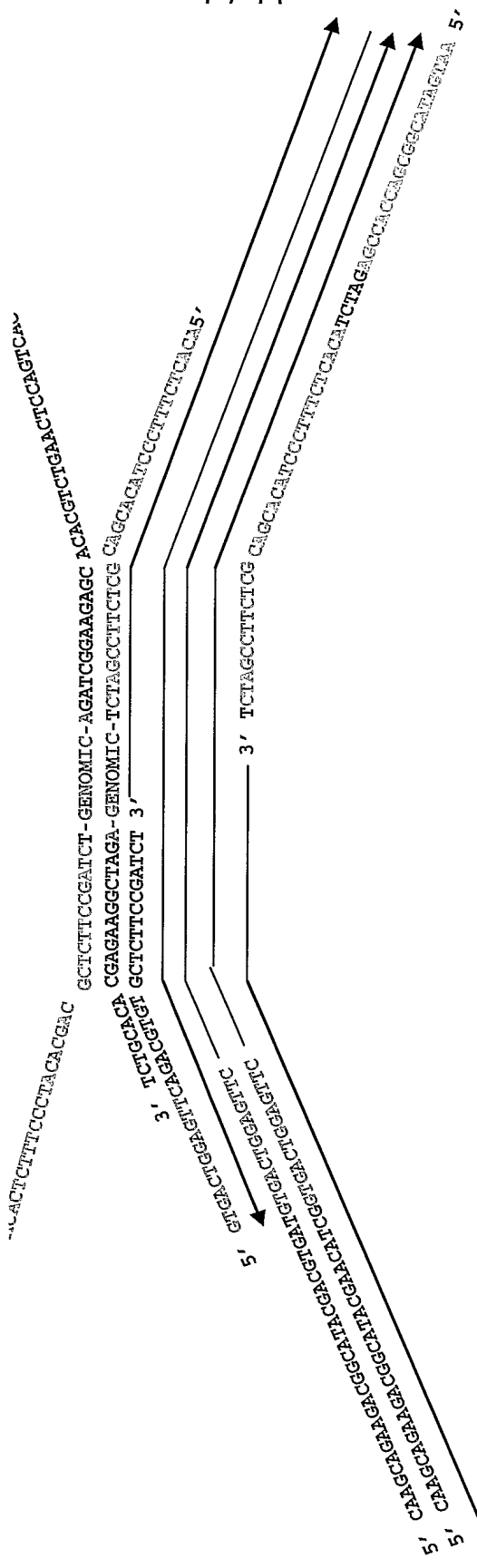


FIG. 6b

IDT Exhibit 1018.0299

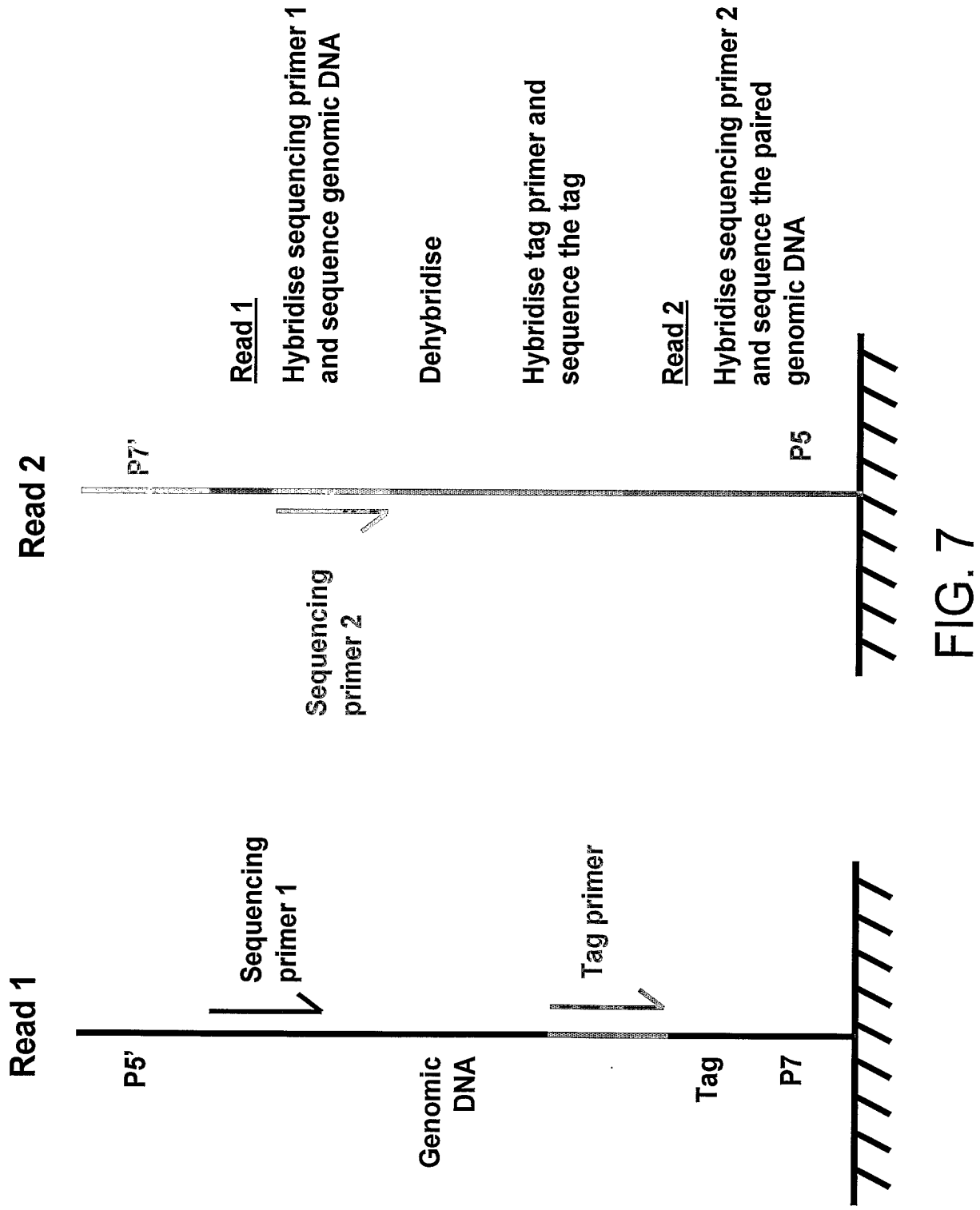


FIG. 7

IDT Exhibit 1018.0300

Tag	Genome	Clus No	BAC R1	BAC R2	PhiX R1	PhiX R2
ATCACG	PhiX	6522	0%	0%	96%	92%
CGATGT	PhiX	3548	0%	0%	96%	92%
TTAGGC	PhiX	10394	0%	0%	97%	93%
TGACCA	PhiX	6842	0%	0%	97%	93%
ACAGTG	PhiX	6677	0%	0%	97%	93%
CAGATC	PhiX	11231	0%	0%	96%	93%
GCCAAT	BAC	6174	97%	92%	0%	0%
ACTTGA	BAC	5375	97%	93%	0%	0%
GATCAG	BAC	5404	97%	93%	0%	0%
TAGCTT	BAC	10239	97%	94%	0%	0%
GGCTAC	BAC	9734	97%	93%	0%	0%
CTTGTA	BAC	9704	97%	93%	0%	0%

Expected Tag	Observed Tag	Genome	Clus No	BAC R1	BAC R2	PhiX R1	PhiX R2
TGACCA	TGACAA	PhiX	8	0%	0%	100%	88%
GCCAAT	TCCAAT	BAC	34	97%	94%	0%	0%
GATCAG	TATCAG	BAC	28	100%	96%	0%	0%

FIG. 8

IDT Exhibit 1018.0301

Forked Adaptor Primers SBS3/SBS12 (indexing/paired ends)

Adaptor oligo 2 5′ ACACTCTTTCCCTACACGACGCTCTTCCGATCT 3′

Adaptor oligo 1 (short) 5′ P-GATCGGAAGAGCACACGTCT 3′

5′ ACACTCTTTCCCTACACGAC GCTCTTCCGATCT 3′

3′ TCTGCACACGAGAAGGCTAG 5′

PCR Primer		
Universal LPX primer	5′	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCxT 3′
Lpx primer	5′	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCxT
Tag1 primer (CGTGAT)	5′	CAAGCAGAAGACGGCATACGAGATCGTGATGTGACTGGAGTTC 3′
Tag2 primer (ACATCG)	5′	CAAGCAGAAGACGGCATACGAGATACATCGGTGACTGGAGTTC 3′
Tag3 primer (GCCTAA)	5′	CAAGCAGAAGACGGCATACGAGATGCCTAAGTGACTGGAGTTC 3′
Tag4 primer (TGGTCA)	5′	CAAGCAGAAGACGGCATACGAGATTGGTCAGTGACTGGAGTTC 3′
Tag5 primer (CACTGT)	5′	CAAGCAGAAGACGGCATACGAGATCACTGTGTGACTGGAGTTC 3′
Tag6 primer (ATTGGC)	5′	CAAGCAGAAGACGGCATACGAGATATTGGCGTGACTGGAGTTC 3′
Tag7 primer (GATCTG)	5′	CAAGCAGAAGACGGCATACGAGATGATCTGGTGACTGGAGTTC 3′
Tag8 primer (TCAAGT)	5′	CAAGCAGAAGACGGCATACGAGATTCAAGTGTGACTGGAGTTC 3′
Tag9 primer (CTGATC)	5′	CAAGCAGAAGACGGCATACGAGATCTGATCGTGACTGGAGTTC 3′
Tag10 primer (AAGCTA)	5′	CAAGCAGAAGACGGCATACGAGATAAGCTAGTGACTGGAGTTC 3′
Tag11 primer (GTAGCC)	5′	CAAGCAGAAGACGGCATACGAGATGTAGCCGTGACTGGAGTTC 3′
Tag12 primer (TACAAG)	5′	CAAGCAGAAGACGGCATACGAGATTACAAGGTGACTGGAGTTC 3′

Sequencing primers		
SBS12	5′	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATC 3′
SBS12rev	5′	GATCGGAAGAGCACACGTCTGAACTCCAGTCAC 3′
SBS3	5′	ACACTCTTTCCCTACACGACGCTCTTCCGATC 3′

FIG. 9

IDT Exhibit 1018.0302

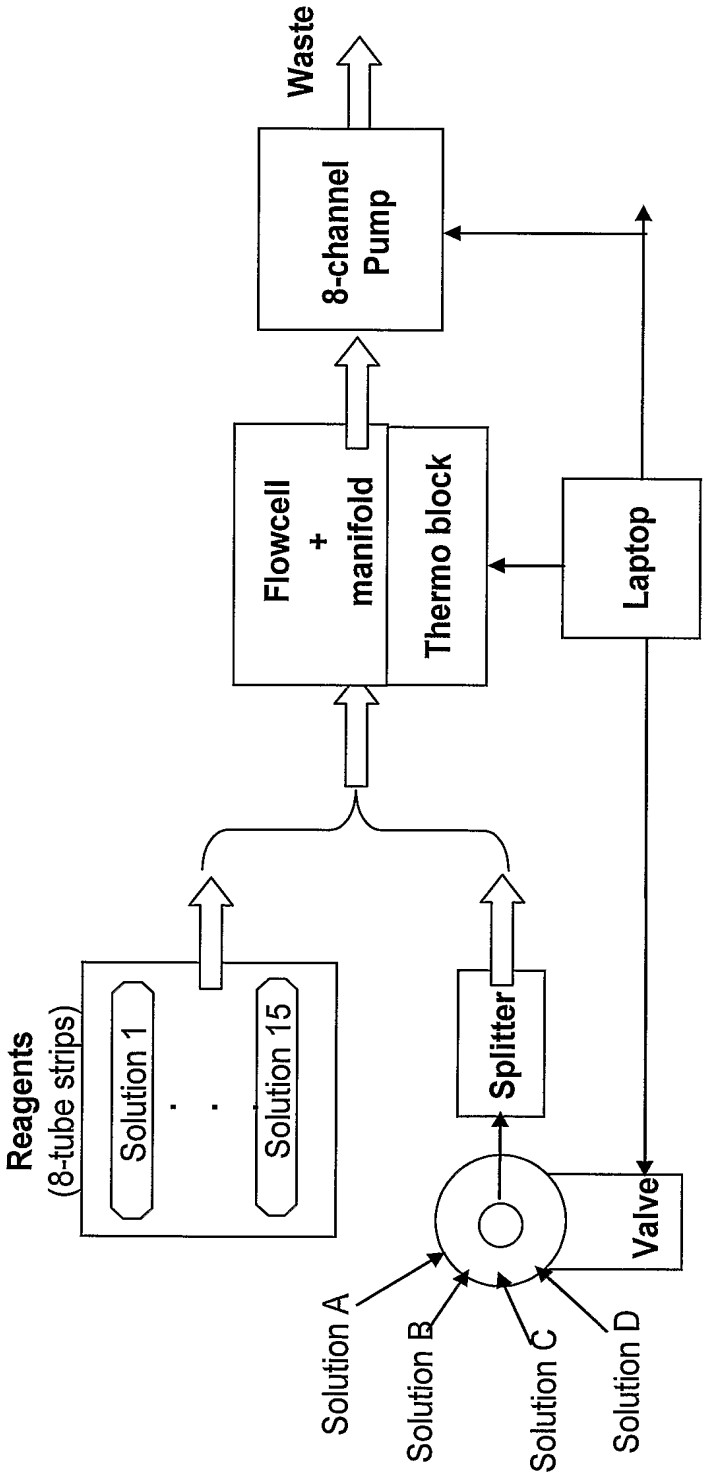


FIG. 10

IDT Exhibit 1018.0303

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

In Re Application of:	AMORESE, Douglas, et al.	Confirmation No.:	7076
Application No.:	14/540,917	Group Art Unit:	1637
Filed:	November 13, 2014	Examiner:	STRZELECKA, Teresa E.

FOR: **COMPOSITIONS AND METHODS FOR IDENTIFICATION OF A DUPLICATE SEQUENCING READ**

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

INFORMATION DISCLOSURE STATEMENT UNDER 37 C.F.R. § 1.97(b)

In accordance with the duty of disclosure set forth in 37 C.F.R. § 1.56, Applicant(s) hereby submits the following information in conformance with 37 C.F.R. §§1.97 and 1.98.

☒ Pursuant to 37 C.F.R. § 1.98, a copy of each non-US patent document cited in the attached Form PTO/SB/08 is enclosed.

☐ No copies of the publications listed on the attached Form PTO/SB/08 are being provided pursuant to 37 C.F.R. § 1.98(d) because the publications were previously cited by or submitted to the Office in prior Application Serial No. _____ to which the above-identified application claims priority under 35 U.S.C. § 120.

☒ No copies of any U.S. patents or U.S. patent application publications listed on the attached Form PTO/SB/08 are being provided pursuant to 37 C.F.R. § 1.98.

☒ Publication(s) 1-5 listed on the attached Form PTO/SB/08 were cited in an International Search Report in International Application No. PCT/US2014/065530 mailed on March 5, 2015.

☐ Enclosed is a copy of a non-English publication(s) _____ Pursuant to §609 of the M.P.E.P., Applicant submits the attached foreign search or examination report, which cites such non-English language publication(s).

☐ Enclosed is a copy of a non-English publication(s) _____ English language publication _____ (copy enclosed) claims priority from this non-English publication.

☐ Enclosed is an explanation of non-English publication(s) _____ for which an English translation is not available.

☐ Enclosed is an English translation of non-English publication(s) _____ cited in the attached Form PTO/SB/08.

☐ Enclosed is a copy of pending patent Application Serial No. _____.

This Information Disclosure Statement is filed within any one of the following time periods:

☐ within three months from the filing date of this national application other than a CPA under 37 C.F.R. § 1.53(d);

☐ within three months from the date of entry of the national stage as set forth in 37 C.F.R. § 1.491 in this international application;

☒ before the mailing date of a first office action on the merits; or

☐ before the mailing of a first office action after the filing of a request for continued examination under 37 C.F.R. § 1.114.

It is respectfully requested that the Examiner consider the above-noted information and return an initialed copy of the attached Form PTO/SB/08 to the undersigned.

Dated: March 31, 2015

Respectfully submitted,
COOLEY LLP

COOLEY LLP
ATTN: Patent Group
1299 Pennsylvania Avenue NW, Suite 700
Washington, DC 20004

Tel: (202) 728-7138
Fax: (202) 842-7899

By: /Angie Mah/
Angie Mah
Reg. No. 64,902

115283294 v1

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

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
10 May 2012 (10.05.2012)

PCT

(10) International Publication Number
WO 2012/061832 A1

(51) International Patent Classification:
C12N 15/10 (2006.01) *C40B 40/08* (2006.01)
C12N 15/90 (2006.01) *C40B 50/06* (2006.01)

(21) International Application Number:
PCT/US2011/059642

(22) International Filing Date:
7 November 2011 (07.11.2011)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
61/410,671 5 November 2010 (05.11.2010) US
13/025,022 10 February 2011 (10.02.2011) US
13/080,345 5 April 2011 (05.04.2011) US

(71) Applicant *(for all designated States except US)*: **ILLUMINA, INC.** [US/US]; 9885 Towne Centre Drive, San Diego, California 92121 (US).

(72) Inventors; and
(75) Inventors/Applicants *(for US only)*: **STEEMERS, Frank J.** [NL/US]; 9885 Towne Centre Drive, San Diego, California 92121 (US). **GUNDERSON, Kevin** [US/US]; 9885 Towne Centre Drive, San Diego, California 92121 (US). **ROYCE, Thomas** [US/US]; 9885 Towne Centre Drive, San Diego, California 92121 (US). **PIGNATELLI, Natasha** [US/US]; 9885 Towne Centre Drive, San Diego, California 92121 (US). **GORYSHIN, Igor Yu** [US/US]; 726 Post Road, Madison, Wisconsin 53713 (US). **CARUCCIO, Nicholas** [US/US]; 726 Post Road, Madison, Wisconsin 53713 (US). **MAFFITT, Mark** [US/US]; 726 Post Road, Madison, Wisconsin 53713 (US). **JENDRISAK, Jerome** [US/US]; 726 Post Road, Madison, Wisconsin 53713 (US). **AMINI, Sasan** [IR/US]; 9885 Towne Centre Drive, San Diego, California 92121 (US). **KAPER, Fiona** [US/US]; 9885 Towne Centre Drive, San Diego, California 92121 (US). **TURK, Casey** [US/US]; 9885 Towne Centre Drive, San Diego, California 92121 (US). **KAHLOR, Reza** [US/US]; 9885 Towne Centre Drive, San Diego, California 92121 (US).

(74) Agents: **FAN, Calvin** et al.; 9885 Towne Centre Drive, San Diego, California 92121 (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, 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.

[Continued on next page]

(54) Title: LINKING SEQUENCE READS USING PAIRED CODE TAGS

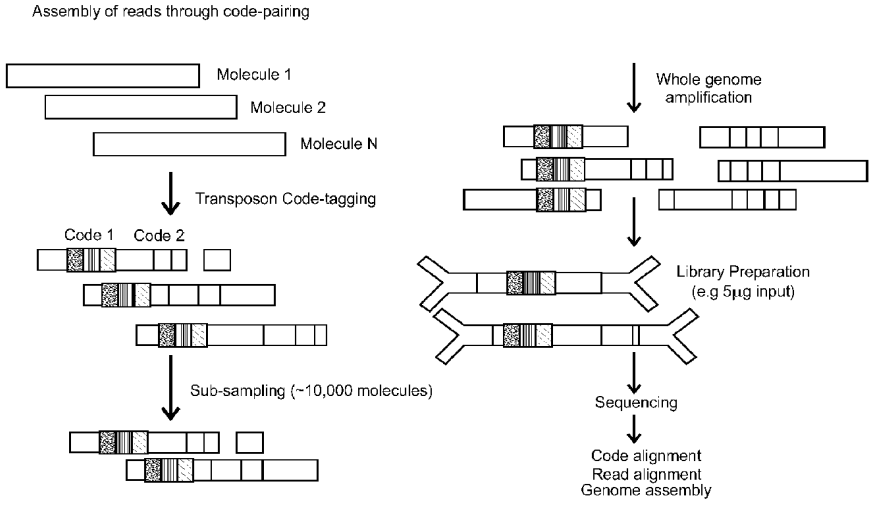


Fig. 31

(57) Abstract: Artificial transposon sequences having code tags and target nucleic acids containing such sequences. Methods for making artificial transposons and for using their properties to analyze target nucleic acids.

WO 2012/061832 A1

(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, MD, 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, ML, MR, NE, SN, TD, TG).

Declarations under Rule 4.17:

— *as to applicant's entitlement to apply for and be granted a patent (Rule 4.17(ii))*

— *as to the applicant's entitlement to claim the priority of the earlier application (Rule 4.17(iii))*

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))*

LINKING SEQUENCE READS USING PAIRED CODE TAGS

RELATED APPLICATIONS

This application is a continuation-in-part of U.S. Application No. 13/080,345, filed April 5, 2011, entitled "LINKING SEQUENCE READS USING PAIRED CODE TAGS," which is a continuation-in-part of U.S. Application No. 13/025,022, filed February 10, 2011 entitled "LINKING SEQUENCE READS USING PAIRED CODE TAGS"; and claims the benefit of priority to U.S. provisional Application No. 61/410,671, filed November 5, 2010, entitled "IMPROVED METHOD FOR MAINTAINING MATE-PAIR DNA SEQUENCING LIBRARY DIVERSITY". The contents of these applications are incorporated herein by reference in their entirety.

FIELD OF THE INVENTION

Embodiments of the present invention relate to the fields of biology and genomics. Some embodiments of the present invention relate to methods and compositions that include certain transposon sequences. Some such methods and compositions include analyzing target nucleic acids.

BACKGROUND OF THE INVENTION

The detection of specific nucleic acid sequences present in a biological sample has been used, for example, as a method for identifying and classifying microorganisms, diagnosing infectious diseases, detecting and characterizing genetic abnormalities, identifying genetic changes associated with cancer, studying genetic susceptibility to disease, and measuring response to various types of treatment. A common technique for detecting specific nucleic acid sequences in a biological sample is nucleic acid sequencing.

Nucleic acid sequencing methodology has evolved significantly from the chemical degradation methods used by Maxam and Gilbert and the strand elongation methods used by Sanger. Today several sequencing methodologies are in use which allow for the

IDT Exhibit 1018.0308

parallel processing of nucleic acids all in a single sequencing run. As such, the information generated from a single sequencing run can be enormous.

SUMMARY OF THE INVENTION

Some embodiments of the present invention relate to methods and compositions that include certain transposon sequences. Some such methods and compositions include analyzing target nucleic acids. Some embodiments include methods of preparing a template nucleic acid. Some such methods include: (a) providing a target nucleic acid; (b) providing a plurality of transposon sequences, each transposon sequence comprising a first transposase recognition site, a second transposase recognition site having a barcode disposed therebetween; and (c) contacting the target nucleic acid with the plurality of transposon sequences under conditions such that at least a portion of said plurality of sequences inserts into the target nucleic acid ("tagementation"), thereby preparing a template nucleic acid.

In some embodiments, the barcode of each transposon is different.

In some embodiments, the barcode comprises a first barcode sequence, a second barcode sequence, said first and second barcode sequences being separated by a linker disposed therebetween.

In some embodiments, the barcode comprises a double-stranded nucleic acid sequence comprising a first strand barcode and a second strand barcode. In some embodiments, the first strand barcode and second strand barcode comprise complementary sequences. In some embodiments, the first strand barcode and second strand barcode comprise non-complementary sequences.

In some embodiments, the linker comprises a nucleic acid.

In some embodiments, the linker comprises a sequencing adapter comprising a first primer site.

In some embodiments, the linker comprises a fragmentation site. In some embodiments, the fragmentation site comprises a first nickase recognition sequence and a second nickase recognition sequence, wherein the cut site for each recognition sequence is the same site. In some embodiments, the fragmentation site comprises a restriction endonuclease recognition sequence.

In some embodiments, the linker comprises a sequencing adapter comprising a first primer site and a fragmentation site. In some embodiments, the linker comprises a

IDT Exhibit 1018.0309

sequencing adapter comprising a first primer site and a second primer site having a fragmentation site disposed therebetween.

Some embodiments include methods of preparing a library of template nucleic acids. Some such methods include: (a) providing a target nucleic acid; (b) providing a plurality of transposon sequences, each transposon sequence comprising a first transposase recognition site and a second transposase recognition site having a barcode disposed therebetween, wherein the barcode comprises a first barcode sequence and a second barcode sequence being separated by a sequencing adapter disposed therebetween, wherein the sequencing adapter comprises a first primer site and a second primer site; and (c) contacting the target nucleic acid with the plurality of transposon sequences under conditions such that at least a portion of said plurality of transposon sequences inserts into the target nucleic acid; and (d) amplifying at least a portion of the target nucleic acid by hybridizing a primer to said first primer site and a primer to said second primer site, such that the amplification product comprises a first barcode sequence and a second barcode sequence, thereby preparing a library of template nucleic acids.

In some embodiments, the barcode of each transposon is different.

Some methods also include a step subsequent to (c) and prior to (d) comprising reducing the number of target nucleic acid molecules comprising inserted transposon sequences.

In some embodiments, the sequencing adapter comprises a first primer site, a second primer site, and a non-amplifiable site therebetween. In some embodiments, the non-amplifiable site comprises a nucleic acid. In some embodiments, the non-amplifiable site comprises at least one nucleotide analogue. In some embodiments, the nucleotide analogue does not significantly base-pair with A, C, G or T.

Some embodiments include methods of preparing a library of template nucleic acids. Some such methods include: (a) providing a target nucleic acid; (b) providing a plurality of transposon sequences, each transposon sequence comprising a first transposase recognition site and a second transposase recognition site having a barcode disposed therebetween, wherein the barcode comprises a first barcode sequence and a second barcode sequence being separated by a linker disposed therebetween, wherein the linker comprises a fragmentation site; (c) contacting the target nucleic acid with the plurality of transposon sequences under conditions such that at least a portion of said plurality of transposon sequences inserts into the target nucleic acid; and (d) fragmenting

IDT Exhibit 1018.0310

said target nucleic acid at said fragmentation sites, such that at least a portion of the fragmented nucleic acids each comprise a first barcode sequence and a second barcode sequence, thereby preparing a library of template nucleic acids.

In some embodiments, the barcode of each transposon is different.

In some embodiments, the linker comprises a nucleic acid.

In some embodiments, the fragmentation site comprises a first nickase recognition sequence, a second nickase recognition sequence, wherein the cut site for each recognition sequence is the same site.

In some embodiments, the fragmentation site comprises a restriction endonuclease recognition sequence.

In some methods, the fragmenting comprises contacting the target nucleic acid with a polymerase. Useful polymerases include those with exonuclease activity (such as 3' to 5' activity, *e.g., E. coli* DNA polymerase III, or 5' to 3' activity, *e.g., E. coli* DNA polymerase I), non-strand-displacing (e.g. T4 DNA polymerase) and strand-displacing activities (e.g. *Bst* DNA polymerase, large fragment).

Some methods also include ligating a first primer site to a first end of at least one fragmented nucleic acid. Some methods also include ligating a second primer site to the second end of the at least one fragmented nucleic acid.

Some methods also include amplifying said at least one fragmented nucleic acid by hybridizing a primer to the first primer site and a primer to the second primer site.

In some embodiments, the linker further comprises a first primer site.

In some embodiments, the linker comprises a sequencing adapter comprises a first primer site, a second primer site having the fragmentation site disposed therebetween. Some methods also include amplifying said at least a portion of the fragmented nucleic acids by hybridizing a primer to said first primer site and a primer to said second primer site.

In some embodiments, the target nucleic acid comprises genomic DNA.

In some embodiments, the first transposase recognition site comprises a mosaic element.

In some embodiments, the sequences of the first barcode and the second barcode comprise the reverse complements of each other.

Some embodiments include libraries of template nucleic acids prepared by any one of the foregoing methods.

IDT Exhibit 1018.0311

Some embodiments include methods of preparing a transposon sequence. Some such methods include: (a) providing a transposon template nucleic acid sequence comprising a first transposase recognition site, a first barcode 3′ thereof, and a linker 3′ thereof; and (b) extending the transposon template sequence by hair-pin extension, such that the extended transposon template sequence further comprises sequences complementary to said barcode and to said transposase recognition site, thereby preparing a transposon comprising a first transposase recognition site, a first barcode, a linker, a second barcode, and a second transposase recognition site.

In some embodiments, the barcode comprises a random sequence.

In some embodiments, the transposase recognition site comprises a mosaic element.

In some embodiments, the barcode comprises at least about five nucleotides.

In some embodiments, the linker comprises a nucleic acid.

In some embodiments, the linker comprises a fragmentation site. In some embodiments, the fragmentation site comprises a first nickase recognition sequence, a second nickase recognition sequence, wherein the cut site for each recognition sequence is the same site. In some embodiments, the fragmentation site comprises a restriction endonuclease recognition sequence.

In some embodiments, the linker comprises a sequencing adapter. In some embodiments, a sequencing adapter comprises a first primer site and second primer site. In some embodiments, the sequencing adapter comprises a first primer site and a second primer site having a fragmentation site disposed therebetween. In some embodiments, the sequencing adapter comprises a first primer site and a second primer site, having a non-amplifiable site disposed therebetween.

In some embodiments, the non-amplifiable site comprises a nucleic acid. In some embodiments, the non-amplifiable site comprises at least one nucleotide analogue. In some embodiments, the nucleotide analogue does not significantly base-pair with A, C, G or T.

Some embodiments include methods of preparing a plurality of transposon sequences. Some such method include repeating steps (a) and (b) of any one of the foregoing methods of preparing a transposon sequence, wherein the barcode of each transposon template nucleic acid is different.

IDT Exhibit 1018.0312

Some embodiments include transposons prepared by the method of any one of the foregoing methods of preparing a transposon sequence.

Some embodiments include methods of analyzing a target nucleic acid. Some such methods include: (a) providing a template nucleic acid, wherein the template nucleic acid comprises the target nucleic acid and a plurality of markers inserted therein, wherein each marker sequence comprises a barcode comprising a first barcode sequence and a second barcode sequence, having a sequencing adapter disposed therebetween, wherein the sequencing adapter comprises a first primer site; (b) obtaining sequence data from said template nucleic acid; and (c) assembling a representation of at least a portion of said target nucleic acid from said sequence data.

In some embodiments, the barcode of each marker is different.

In some embodiments, the barcode comprises a double-stranded nucleic acid sequence comprising a first strand barcode and a second strand barcode. In some embodiments, the first strand barcode and second strand barcode comprise complementary sequences. In some embodiments, the first strand barcode and second strand barcode comprise non-complementary sequences.

In some embodiments, obtaining sequence data comprises hybridizing a primer to said first primer site, and extending said primer.

In some embodiments, the sequencing adapter comprises a second primer site, and said obtaining sequence data comprises hybridizing a primer to the second primer site and extending said primer. In some embodiments, the primers hybridize to the sites in opposite orientations.

In some embodiments, the sequences of the first barcode sequence and the second barcode sequence comprise the reverse complements of each other.

In some embodiments, the assembling step comprises identifying more than one sequencing read comprising the first barcode sequence or second barcode sequence of a barcode.

In some embodiments, the presence of a first barcode sequence or second barcode sequence of a barcode in more than one sequencing read is indicative of the more than one sequencing reads representing sequences adjacent to each other in the target nucleic acid.

In some embodiments, each marker sequence comprises a first host tag and second host tag having the barcode disposed therebetween. In some embodiments, the

IDT Exhibit 1018.0313

first and second host tag of a marker comprises the same sequence. In some embodiments, the assembling step further comprises identifying more than one sequencing read comprising the same host tags.

In some embodiments, the target nucleic acid comprises genomic DNA.

Other embodiments include artificial transposon sequences. Some such artificial transposon sequences include: a first transposase recognition site and a second transposase recognition site having a barcode disposed therebetween, wherein the barcode comprises a first barcode sequence and a second barcode sequence, said first and second barcode sequences being separated by a linker.

In some embodiments, the barcode comprises a double-stranded nucleic acid sequence comprising a first strand barcode and a second strand barcode. In some embodiments, the first strand barcode and second strand barcode comprise complementary sequences. In some embodiments, the first strand barcode and second strand barcode comprise non-complementary sequences.

In some embodiments, the first transposon recognition site comprises a mosaic element.

In some embodiments, the linker comprises a nucleic acid.

In some embodiments, the linker comprises a fragmentation site. In some embodiments, the fragmentation site comprises a first nickase recognition sequence and a second nickase recognition sequence, wherein the cut site for each recognition sequence is the same site. In some embodiments, the fragmentation site comprises a restriction endonuclease recognition sequence.

In some embodiments, the linker comprises a sequencing adapter comprising a first primer site. In some embodiments, the sequencing adapter comprises a second primer site. In some embodiments, the sequencing adapter comprises a first primer site and a second primer site, having a non-amplifiable site disposed therebetween. In some embodiments, the non-amplifiable site comprises a nucleic acid. In some embodiments, the non-amplifiable site comprises at least one nucleotide analogue. In some embodiments, the nucleotide analogue does not significantly base-pair with A, C, G or T.

In some embodiments, the linker comprises a sequencing adapter comprising a first primer site and a second primer site, having a fragmentation site disposed therebetween.

In some embodiments, a sequencing primer is hybridized to said first primer site.

IDT Exhibit 1018.0314

In some embodiments, a sequencing primer is hybridized to said first transposase recognition site.

Some embodiments include populations of artificial transposon sequences comprising a plurality of any one of the foregoing artificial transposon sequences.

Some embodiments include target nucleic acids or copies thereof having a population of the artificial transposon sequences integrated therein, wherein the artificial transposon sequences comprise any one of the artificial transposon sequences described herein.

Some embodiments include populations of genomic DNA fragments or copies thereof having a population of the artificial transposon sequences integrated therein, wherein the artificial transposon sequences comprise any one of the artificial transposon sequences described herein.

Some embodiments include genomes having a population of artificial transposon sequences integrated therein, wherein the artificial transposon sequences comprise any one of the artificial transposon sequences described herein.

Some embodiments include isolated template nucleic acids. Some such isolated template nucleic acids include at least a portion of a target nucleic acid or copy thereof and at least two markers inserted therein, wherein each marker sequence comprises a barcode.

In some embodiments, the barcode of each marker is different.

In some embodiments, the barcode comprises a double-stranded nucleic acid sequence comprising a first strand barcode and a second strand barcode. In some embodiments, the first strand barcode and second strand barcode comprise complementary sequences. In some embodiments, the first strand barcode and second strand barcode comprise non-complementary sequences.

Some embodiments also include a transposase recognition site.

Some embodiments also include a first transposase recognition site and a second transposes recognition site, having the at least a portion of a target nucleic acid or copy thereof disposed therebetween.

Some embodiments also include a host tag. Some embodiments also include a first host tag and a second host tag, having the at least a portion of a target nucleic acid or copy thereof disposed therebetween. Some embodiments also include a first host tag and a second host tag, having a transposase recognition site disposed therebetween, wherein

IDT Exhibit 1018.0315

the transposase recognition site comprises a first transposase recognition site and a second transposase recognition site, having the at least a portion of a target nucleic acid or copy thereof disposed therebetween.

In some embodiments, the host tag comprises nine nucleotides.

In some embodiments, the target nucleic acid comprises genomic DNA.

Some embodiments include a plurality of template nucleic acids comprising the template nucleic acid of any one of the isolated template nucleic acids described herein, wherein a first template nucleic acid of said plurality comprises a first barcode, and a second template nucleic acid comprises a second barcode, wherein the first and second barcodes are indicative of template nucleic acid sequences being adjacent to one another in a sequence representation of the target nucleic acid. In some embodiments, the first barcode comprises the reverse complement sequence of the second barcode.

Some embodiments include substrates comprising nucleotide sequences attached thereto, at least one of said nucleotide sequences comprising any one of the template nucleic acids described herein, or any one of the plurality of template nucleic acids described herein. In some embodiments, the substrate can include a solid support selected from the group consisting of spheres, microparticles, beads, membranes, slides, plates, micromachined chips, tubes, microwells, microfluidic devices, channels, and filters.

BRIEF DESCRIPTION OF THE DRAWINGS

FIG. 1 depicts a schematic a transposon sequence. M indicates a first transposon recognition sequence. C1 indicates a first barcode sequence. The linker sequence is indicated by a rectangle with a double border. C1′ indicates a second barcode sequence. M' indicates a second transposase recognition sequence. Exemplary linker sequences are also shown. For example, linker sequences can have fragmentation sites, such as those with nickase recognition sites (e.g. N1, N2) or a restriction endonuclease recognition sequence (indicated by RE). In other embodiments, linkers can contain one or more primer sites (e.g. A, B), corresponding to complementary primers A′ and B′. A non-amplifiable site is indicated by nA.

FIG. 2 depicts a double-stranded transposon sequence comprising transposon recognition sites (TRS and TRS′), a fragmentation site, and a barcode comprising four barcode sequences (CODE 1, CODE 2, CODE 3, and CODE 4). In the embodiment

IDT Exhibit 1018.0316

shown, CODE 1 and CODE 3, and CODE 2 and CODE 4 are non-complementary sequences.

FIG. 3 illustrates single-stranded transposons having bicodes or unicodes.

FIG. 4 illustrates preparation of transposome having different transposase recognition sequences.

FIG. 5 depicts a transposome complex that has one transposon sequence and two transposases.

FIG. 6 illustrates gap fill-in and ligation reactions.

FIG. 7 depicts "looped" complexes.

FIG. 8 depicts a complex with one transposon and two transposase subunits ensures that a known combination of codes is inserted into target DNA, without fragmenting the target DNA. This configuration allows linking DNA fragments after fragmentation using the code N-Code N' combination.

FIG. 9 illustrates strand-displacement synthesis to copy identical flanking genomic DNA sequences (g-codes).

FIG. 10 illustrates nicking followed by 3'-exo digestion to create gaps in genomic DNA.

FIG. 11 illustrates transposons with engineered bicodes and random bi-codes.

FIG. 12. Transposon-transposase complex. Complexes are formed from two half-transposons, each containing one ME region and a random or specific code region. Upon insertion, the transposase enzyme binds to both transposons preventing DNA fragmentation. The ends of the transposons have "sticky ends" allowing ligation of the two ends. A specific endonuclease fragmentation system nicks each strand. Denaturation creates DNA fragments in which each DNA fragment contains a linked code with the neighboring fragment.

FIG. 13. Transposons with amplifiable and non-amplifiable linkers between the ME sequences.

FIG. 14 depicts use of bi-code junction-tagging.

FIG. 15 depicts a library preparation method that involves ligation of a forked adapter sequence. Steps include: library preparation; extension and A-tailing; adapter ligation; and PCR amplify, cluster, and sequence products.

FIG. 16 depicts an exemplary scheme to prepare a transposon sequence with reverse complements

IDT Exhibit 1018.0317

FIG. 17 depicts a method of preparing a transposon sequence which includes the use of tailed-oligonucleotides comprising mosaic elements (ME). The transposon sequence includes primer sites P3 and P4, a barcode comprising Code X and Code Y, and a fragmentation site.

FIG. 18 depicts the integration of an exemplary Tn5 transposon comprising transposase recognition sequences (indicated by mosaic elements ME) into a target nucleic acid. The insertion results in duplication of the integration site, shown as single-stranded A′B′C′D′E′F′G′H′I′ and single-stranded ABCDEFGHI. The figure also shows an optional repair step to fill in the single-stranded region. In some embodiments, the filled-in region can be used as a host tag.

FIG. 19 shows a tagmentation method using transposomes that have symmetrical transposable end sequences.

FIG. 20 depicts an exemplary embodiment of a contacting a target nucleic acid with a library of transposons, which integrates into the target nucleic acid.

FIG. 21 depicts an optional step of fragmenting a linker into two parts, or an optional step of virtually fragmenting the linker by amplification using a primer A′ and/or a primer B′. Fragmentation of the linker generates a library of template nucleic acids. In this figure, the filled-in region is indicated by a thick line.

FIG. 22 depicts an examplary embodiment using tailed-oligonucleotides to obtain a population of template nucleic acids comprising a first universal primer site (A) and a second universal primer site (B). In step 1, tailed-oligonucleotides comprising universal primer site sequences anneal to a template nucleic acid comprising primer site sequences (P and P′) and are extended. In step 2, the tailed oligonucleotides of step 1 may anneal to the products of step 1 and be extended. Alternatively, step 2 can include an amplification step with the inclusion of oligonucleotides (*e.g.,* oligonucleotides comprising A′ and B′ sequences). The products of step 2 are depicted in step 3.

FIG. 23 illustrates an embodiment for code-tagging individual molecules with unique tags, allowing normalization of the original molecule representation, even after amplification.

FIG. 24 depicts how linked transposon ends result in fragmented DNA, simultaneously linking the top and bottom strand of the target DNA.

FIG. 25. Targeted Tagmentation using ssDNA. (1) Locus-specific oligo probes are annealed to regions of interest in the denatured DNA sample. An optional "gap-fill"

IDT Exhibit 1018.0318

extension (polymerase, ligase, nucleotides, and buffer) can be employed to create dsDNA between sites of probe annealing. (2) Tagmentation mix (i.e. Tn5-mosaic end transposon complexes and appropriate reaction buffer) is added to the annealed DNA. (3) Tagmentation only occurs at regions of dsDNA leading to "targeted" tagmentation (and nicking) at sites where the oligo probes have annealed. (4) Strand-displacement synthesis resolves the nicks into amplifiable library elements flanked by the universal ME primer. This ME primer can be used in a subsequent PCR step to attach P5-Seq1 and P7-Seq2 sequencing tails to the tagments.

FIG. 26. Targeted Tagmentation using dsDNA and Strand Invasion/Triplex formation. D-loop formation via strand invasion using recA coated oligonucleotides or LNA probes can be used to create D-loops and target transposase complexes to loci or regions of interest (ROI) in dsDNA. Targeted transposition can also be accomplished by triplex formation using pyrimidine-rich oligonucleotides. The targeting probe or oligonucleotide is conjugated to the mosaic elements (ME) on the transposase-transposon complex. (1) Strand invasion opens a D-loop in the dsDNA bringing the transposase complex into the vicinity of the targeted region. Ideally two such probes will be used to generate libraries spanning the region between the probes. (2) Localization of the transposon complex leads to insertion in the vicinity of the D-loop or triplex structure. (3) Strand-displacement is use to finish the library element flanked by universal ME primers. PCR can be used append sequencing primers using the ME ends.

FIG. 27 depicts tethering of a circularized intermediate to the surface of a particle.

FIG. 28 depicts a schematic representation of using paired bar codes (indicated by Code 1, Code 2, Code 5) for code-pair sequencing of template contig sequences to assemble the sequence of an original target nucleic acid.

FIG. 29 depicts an exemplary embodiment of a method including fragmentation.

FIG. 30 depicts an exemplary embodiment of a method that includes optional steps of amplifying a target nucleic acid by PCR or whole genome amplification (WGA). In this figure, the filled-in region is indicated by N_1 and N_2.

FIG. 31 depicts an exemplary embodiment of a method for assembling short sequencing reads using a linked read strategy. Optional steps of sub-sampling, code alignment, read assembly, and genome assembly are also illustrated.

FIG. 32 illustrates the preparation of nested sequencing libraries.

IDT Exhibit 1018.0319

FIG. 33 illustrates a two-step tagging method using different transposomes.

FIG. 34 illustrates single cell/DNA tagmentation in microwells.

FIG. 35 illustrates single cell/DNA tagmentation in droplets.

FIG. 36 depicts a graph showing that the proportion of template nucleic acids that are useful can decrease as the average distance between sites of integration increases.

FIG. 37 depicts an embodiment using restriction endonucleases to generate randomer sticky ends.

DETAILED DESCRIPTION

Some embodiments of the present invention relate to methods and compositions that include certain transposon sequences. Some such methods and compositions include analyzing target nucleic acids. Generally, methods of analyzing nucleic acids include preparing a library of template nucleic acids of a target nucleic acid, obtaining sequence data from the library of template nucleic acids, and assembling a sequence representation of the target nucleic acid from such sequence data. However, assembly of a sequence representation using traditional methods is met with several challenges. For example, sequencing data is obtained using short reads and thus there is difficulty assembling contiguous reads through repetitive sequences of a target nucleic acid. In addition, many contigs are required to be constructed to represent a genome such as the human genome. Methods such as paired-end sequencing can be used to mitigate the difficulties of assembling sequence data from many short reads. However, the paired-end sequencing methods require longer template nucleic acids to span repetitive sequences in a target nucleic acid; preparation of template nucleic acids is inefficient and thus requires larger amounts of DNA; the number of different paired-ends is limited; and a reference genome is required to verify any resulting sequence representation.

Some nucleic acids of interest, such as genomic DNAs, comprise long molecules with ordered sequence. Methods to sequence such molecules tend to be highly parallelized and include sequencing short libraries in order to create high throughput. Consequently, such methods require assembly of the shorter reads to obtain ordering information, *i.e.,* a sequence representation of the target nucleic acid. Shotgun sequencing approaches uses DNA fragmentation; however, order information is lost during fragmentation. Advantageously, some methods and compositions provided herein can be used to obtain shorter reads of template nucleic acids in which the order

IDT Exhibit 1018.0320

information is preserved. Thus, assembly of shorter reads can be performed without the requirement of a reference genome.

In an exemplary embodiment, a library of template nucleic acids is prepared from a target nucleic acid. The library is prepared by inserting a plurality of unique barcodes throughout the target nucleic acid.

As will be understood, although a barcode is frequently depicted in the figures as a double-stranded, annealed structure, a useful barcode can be a double-stranded region where the two strands are not complementary or not annealed. For example, two noncomplementary regions can be linked informatically, even though they are not complementary in a conventional Watson-Crick base-pairing.

Each barcode includes a first barcode sequence and a second barcode sequence, having a fragmentation site disposed therebetween. The first barcode sequence and second barcode sequences can be identified or designated to be paired with one another. The pairing can be informatic so that a first barcode is associated with a second barcode. The pairing can also be physically associated to form a junction between the two barcodes. For example, the first barcode sequence and second barcode sequences can be the tandem or reverse complements of each other. The target nucleic acid can be fragmented at the fragmentation sites, and a library of template nucleic acids can be prepared from the fragments. Sequencing information can be obtained from the library of template nucleic acids. Advantageously, the paired barcode sequences can be used to assemble sequencing data from the library of template nucleic acids. For example, identifying a first template nucleic acid comprising a first barcode sequence and a second template nucleic acid comprising a second barcode sequence that is paired with the first indicates that the first and second template nucleic acids represent sequences adjacent to one another in a sequence representation of the target nucleic acid. In essence, two sequences originally adjacent in the target nucleic acid may be separated from each other, introducing one barcode of a barcode pair at their points of separation, so that even when sequenced separately, the detection of the paired barcodes indicates the original proximity of the two sequences. Despite having been separated physically, they can be reunited bioinformatically, much as each divided half sought its other half in Plato's *Symposium*. Thus, a sequence representation of the target nucleic acid can be assembled by identifying further sequencing reads comprising paired barcode sequences. Such methods can be used to assemble a sequence representation of a target nucleic acid *de novo*, without the

IDT Exhibit 1018.0321

requirement of a reference genome. Moreover, such methods are also useful to sequence target nucleic acids comprising highly repetitive sequences. The methods can also be used to reassemble alternative isoforms and splice junctions in cDNAs, as well as reassemble single-molecule haplotypes.

Because the methods herein involve addition of one or more tags (such as an identifier code) at the location that previously existed between two adjacent sequences, these methods can be referred to as "junction tagging". More generally, the addition of an identifier tag to an individual nucleic acid can be termed "code tagging".

Definitions

As used herein the term "nucleic acid" and/or "oligonucleotide" and/or grammatical equivalents thereof can refer to at least two nucleotide monomers linked together. A nucleic acid can generally contain phosphodiester bonds; however, in some embodiments, nucleic acid analogs may have other types of backbones, comprising, for example, phosphoramide (Beaucage, *et al., Tetrahedron*, 49:1925 (1993); Letsinger, *J. Org. Chem.,* 35:3800 (1970); Sprinzl, *et al., Eur. J. Biochem.,* 81:579 (1977); Letsinger, *et al., Nucl. Acids Res.,* 14:3487 (1986); Sawai, *et al.*, Chem. Lett., 805 (1984), Letsinger, *et al., J. Am. Chem. Soc.,* 110:4470 (1988); and Pauwels, *et al., Chemica Scripta,* 26:141 (1986)), phosphorothioate (Mag, *et al., Nucleic Acids Res.*, 19:1437 (1991); and U.S. Pat. No. 5,644,048), phosphorodithioate (Briu, *et al., J. Am. Chem. Soc.,* 111:2321 (1989), O-methylphosphoroamidite linkages (see Eckstein, *Oligonucleotides and Analogues: A Practical Approach*, Oxford University Press), and peptide nucleic acid backbones and linkages (see Egholm, *J. Am. Chem. Soc.,* 114:1895 (1992); Meier, *et al., Chem. Int. Ed. Engl.,* 31:1008 (1992); Nielsen, *Nature*, 365:566 (1993); Carlsson, *et al., Nature*, 380:207 (1996)).

Other analog nucleic acids include those with positive backbones (Denpcy, *et al., Proc. Natl. Acad. Sci. USA,* 92:6097 (1995)); non-ionic backbones (U.S. Pat. Nos. 5,386,023; 5,637,684; 5,602,240; 5,216,141; and 4,469,863; Kiedrowshi, *et al.,* Angew. Chem. Intl. Ed. English, 30:423 (1991); Letsinger, *et al., J. Am. Chem. Soc.,* 110:4470 (1988); Letsinger, *et al., Nucleosides & Nucleotides*, 13:1597 (1994); Chapters 2 and 3, *ASC Symposium* Series 580, "Carbohydrate Modifications in Antisense Research", Ed. Y. S. Sanghui and P. Dan Cook; Mesmaeker, *et al., Bioorganic & Medicinal Chem. Lett.,* 4:395 (1994); Jeffs, *et al., J. Biomolecular NMR*, 34:17 (1994); *Tetrahedron Lett.,* 37:743

IDT Exhibit 1018.0322

(1996)) and non-ribose (U.S. Patent No. 5,235,033 and No. 5,034,506, and Chapters 6 and 7, *ASC Symposium* Series 580, "Carbohydrate Modifications in Antisense Research", Ed. Y. S. Sanghui and P. Dan Coo). Nucleic acids may also contain one or more carbocyclic sugars (see Jenkins, *et al.,* Chem. Soc. Rev., (1995) pp. 169 176).

Modifications of the ribose-phosphate backbone may be done to facilitate the addition of additional moieties such as labels, or to increase the stability of such molecules under certain conditions. In addition, mixtures of naturally occurring nucleic acids and analogs can be made. Alternatively, mixtures of different nucleic acid analogs, and mixtures of naturally occurring nucleic acids and analogs may be made. The nucleic acids may be single stranded or double stranded, as specified, or contain portions of both double stranded or single stranded sequence. The nucleic acid may be DNA, for example, genomic or cDNA, RNA or a hybrid, from single cells, multiple cells, or from multiple species, as with metagenomic samples, such as from environmental samples. A nucleic acid can contain any combination of deoxyribo- and ribo-nucleotides, and any combination of bases, including uracil, adenine, thymine, cytosine, guanine, inosine, xanthanine, hypoxanthanine, isocytosine, isoguanine, and base analogs such as nitropyrrole (including 3-nitropyrrole) and nitroindole (including 5-nitroindole), etc.

In some embodiments, a nucleic acid can include at least one promiscuous base. Promiscuous bases can base-pair with more than one different type of base. In some embodiments, a promiscuous base can base-pair with at least two different types of bases and no more than three different types of bases. An example of a promiscuous base includes inosine that may pair with adenine, thymine, or cytosine. Other examples include hypoxanthine, 5-nitroindole, acylic 5-nitroindole, 4-nitropyrazole, 4-nitroimidazole and 3-nitropyrrole (Loakes *et al., Nucleic Acid Res.* 22:4039 (1994); Van Aerschot *et al., Nucleic Acid Res.* 23:4363 (1995); Nichols *et al., Nature* 369:492 (1994); Bergstrom *et al., Nucleic Acid Res.* 25:1935 (1997); Loakes *et al.*, Nucleic Acid Res. 23:2361 (1995); Loakes *et al., J. Mol. Biol.* 270:426 (1997); and Fotin *et al., Nucleic Acid Res.* 26:1515 (1998)). Promiscuous bases that can base-pair with at least three, four or more types of bases can also be used.

As used herein, the term "nucleotide analog" and/or grammatical equivalents thereof can refer to synthetic analogs having modified nucleotide base portions, modified pentose portions, and/or modified phosphate portions, and, in the case of polynucleotides, modified internucleotide linkages, as generally described elsewhere (*e.g.,* Scheit,

IDT Exhibit 1018.0323

Nucleotide Analogs, John Wiley, New York, 1980; Englisch, Angew. *Chem. Int. Ed. Engl.* 30:613-29, 1991; Agarwal, *Protocols for Polynucleotides and Analogs*, Humana Press, 1994; and S. Verma and F. Eckstein, *Ann. Rev. Biochem.* 67:99-134, 1998). Generally, modified phosphate portions comprise analogs of phosphate wherein the phosphorous atom is in the +5 oxidation state and one or more of the oxygen atoms is replaced with a non-oxygen moiety, *e.g.,* sulfur. Exemplary phosphate analogs include but are not limited to phosphorothioate, phosphorodithioate, phosphoroselenoate, phosphorodiselenoate, phosphoroanilothioate, phosphoranilidate, phosphoramidate, boronophosphates, including associated counterions, *e.g.,* H^+, NH_4^+, Na^+, if such counterions are present. Example modified nucleotide base portions include but are not limited to 5-methylcytosine (5mC); C-5-propynyl analogs, including but not limited to, C-5 propynyl-C and C-5 propynyl-U; 2,6-diaminopurine, also known as 2-amino adenine or 2-amino-dA); hypoxanthine, pseudouridine, 2-thiopyrimidine, isocytosine (isoC), 5-methyl isoC, and isoguanine (isoG; *see, e.g.,* U.S. Pat. No. 5,432,272). Exemplary modified pentose portions include but are not limited to, locked nucleic acid (LNA) analogs including without limitation Bz-A-LNA, 5-Me-Bz-C-LNA, dmf-G-LNA, and T-LNA (*see, e.g., The Glen Report*, 16(2):5, 2003; Koshkin *et al., Tetrahedron* 54:3607-30, 1998), and 2'- or 3'-modifications where the 2'- or 3'-position is hydrogen, hydroxy, alkoxy (*e.g.,* methoxy, ethoxy, allyloxy, isopropoxy, butoxy, isobutoxy and phenoxy), azido, amino, alkylamino, fluoro, chloro, or bromo. Modified internucleotide linkages include phosphate analogs, analogs having achiral and uncharged intersubunit linkages (*e.g.,* Sterchak, E. P. *et al.,* Organic Chem., 52:4202, 1987), and uncharged morpholino-based polymers having achiral intersubunit linkages (*see, e.g.,* U.S. Pat. No. 5,034,506). Some internucleotide linkage analogs include morpholidate, acetal, and polyamide-linked heterocycles. In one class of nucleotide analogs, known as peptide nucleic acids, including pseudocomplementary peptide nucleic acids ("PNA"), a conventional sugar and internucleotide linkage has been replaced with a 2-aminoethylglycine amide backbone polymer (*see, e.g.,* Nielsen *et al., Science*, 254:1497-1500, 1991; Egholm *et al., J. Am. Chem. Soc.,* 114: 1895-1897 1992; Demidov *et al., Proc. Natl. Acad. Sci.* 99:5953-58, 2002; *Peptide Nucleic Acids: Protocols and Applications*, Nielsen, ed., Horizon Bioscience, 2004).

As used herein, the term “sequencing read” and/or grammatical equivalents thereof can refer to a repetitive process of physical or chemical steps that is carried out to

IDT Exhibit 1018.0324

obtain signals indicative of the order of monomers in a polymer. The signals can be indicative of an order of monomers at single monomer resolution or lower resolution. In particular embodiments, the steps can be initiated on a nucleic acid target and carried out to obtain signals indicative of the order of bases in the nucleic acid target. The process can be carried out to its typical completion, which is usually defined by the point at which signals from the process can no longer distinguish bases of the target with a reasonable level of certainty. If desired, completion can occur earlier, for example, once a desired amount of sequence information has been obtained. A sequencing read can be carried out on a single target nucleic acid molecule or simultaneously on a population of target nucleic acid molecules having the same sequence, or simultaneously on a population of target nucleic acids having different sequences. In some embodiments, a sequencing read is terminated when signals are no longer obtained from one or more target nucleic acid molecules from which signal acquisition was initiated. For example, a sequencing read can be initiated for one or more target nucleic acid molecules that are present on a solid phase substrate and terminated upon removal of the one or more target nucleic acid molecules from the substrate. Sequencing can be terminated by otherwise ceasing detection of the target nucleic acids that were present on the substrate when the sequencing run was initiated.

As used herein, the term "sequencing representation" and/or grammatical equivalents thereof can refer to information that signifies the order and type of monomeric units in the polymer. For example, the information can indicate the order and type of nucleotides in a nucleic acid. The information can be in any of a variety of formats including, for example, a depiction, image, electronic medium, series of symbols, series of numbers, series of letters, series of colors, etc. The information can be at single monomer resolution or at lower resolution, as set forth in further detail below. An exemplary polymer is a nucleic acid, such as DNA or RNA, having nucleotide units. A series of "A," "T," "G," and "C" letters is a well known sequence representation for DNA that can be correlated, at single nucleotide resolution, with the actual sequence of a DNA molecule. Other exemplary polymers are proteins having amino acid units and polysaccharides having saccharide units.

As used herein the term "at least a portion" and/or grammatical equivalents thereof can refer to any fraction of a whole amount. For example, "at least a portion" can refer to at least about 1%, 2%, 3%, 4%, 5%, 6%, 7%, 8%, 9%, 10%, 15%, 20%, 25%, 30%, 35%,

IDT Exhibit 1018.0325

40%, 45%, 50%, 55%, 60%, 65%, 70%, 75%, 80%, 85%, 90%, 95%, 99%, 99.9% or 100% of a whole amount.

Transposomes

A "transposome" is comprised of at least a transposase enzyme and a transposase recognition site. In some such systems, termed "transposomes", the transposase can form a functional complex with a transposon recognition site that is capable of catalyzing a transposition reaction. The transposase or integrase may bind to the transposase recognition site and insert the transposase recognition site into a target nucleic acid in a process sometimes termed "tagmentation". In some such insertion events, one strand of the transposase recognition site may be transferred into the target nucleic acid.

Some embodiments can include the use of a hyperactive Tn5 transposase and a Tn5-type transposase recognition site (Goryshin and Reznikoff, *J. Biol. Chem.*, 273:7367 (1998)), or MuA transposase and a Mu transposase recognition site comprising R1 and R2 end sequences (Mizuuchi, K., *Cell*, 35: 785, 1983; Savilahti, H, *et al., EMBO J.*, 14: 4893, 1995). An exemplary transposase recognition site that forms a complex with a hyperactive Tn5 transposase (*e.g.*, EZ-Tn5™ Transposase, Epicentre Biotechnologies, Madison, Wisconsin) comprises the following 19b transferred strand (sometimes "M" or "ME") and non-transferred strands: 5′ AGATGTGTATAAGAGACAG 3′, (SEQ ID NO:1), 5′ CTGTCT CTTATACACATCT 3′ (SEQ ID NO:2), respectively.

More examples of transposition systems that can be used with certain embodiments provided herein include *Staphylococcus aureus* Tn552 (Colegio *et al., J. Bacteriol.,* 183: 2384-8, 2001; Kirby C *et al., Mol. Microbiol.,* 43: 173-86, 2002), Ty1 (Devine & Boeke, *Nucleic Acids Res.*, 22: 3765-72, 1994 and International Publication WO 95/23875), Transposon Tn7 (Craig, N L, *Science.* 271: 1512, 1996; Craig, N L, Review in: *Curr Top Microbiol Immunol.,* 204:27-48, 1996), Tn/O and IS10 (Kleckner N, *et al., Curr Top Microbiol Immunol.,* 204:49-82, 1996), Mariner transposase (Lampe D J, *et al., EMBO J.,* 15: 5470-9, 1996), Tc1 (Plasterk R H, *Curr. Topics Microbiol. Immunol.,* 204: 125-43, 1996), P Element (Gloor, G B, *Methods Mol. Biol.,* 260: 97-114, 2004), Tn3 (Ichikawa & Ohtsubo, *J Biol. Chem.* 265:18829-32, 1990), bacterial insertion sequences (Ohtsubo & Sekine, *Curr. Top. Microbiol. Immunol.* 204: 1-26, 1996), retroviruses (Brown, *et al., Proc Natl Acad Sci USA*, 86:2525-9, 1989), and retrotransposon of yeast (Boeke & Corces, *Annu Rev Microbiol.* 43:403-34, 1989). More examples include IS5,

IDT Exhibit 1018.0326

Tn10, Tn903, IS911, and engineered versions of transposase family enzymes (Zhang *et al.,* (2009) *PLoS Genet.* 5:e1000689. Epub 2009 Oct 16; Wilson C. et al (2007) *J. Microbiol. Methods* 71:332-5).

Transposon sequences

Some embodiments provided herein include transposon sequences. In some embodiments, a transposon sequence includes at least one transposase recognition site and at least one barcode. In some embodiments, a transposon sequence includes a first transposon recognition site, a second transposon recognition site, and a barcode disposed therebetween. **FIG. 1** depicts a schematic a transposon sequence.

A transposase recognition site can include two complementary nucleic acid sequences, *e.g.,* a double-stranded nucleic acid or a hairpin nucleic acid, that comprise a substrate for a transposase or integrase. An exemplary embodiment of a transposon sequence comprising a double-stranded nucleic acid comprising non-complementary strands is depicted in **FIG. 2**. In some embodiments, a transposon sequence can include at least a portion comprising at least a portion comprising non-complementary sequences.

— single-stranded sequences

In other embodiments, the transposon sequence contains one or more single-stranded portions. Transposons with single-stranded internal sequences can be useful for avoiding the formation of hairpin structures, or when it is undesirable for transposomes to insert into previously transposome-inserted sequences ("self-insertion"), since many transposases will not insert into single-stranded DNA.

Single-stranded transposons with codes can also be prepared by making single-stranded transposons with hair-pinned ends and removing the hair-pin loops by cutting the DNA with endonucleases, such as restriction endonucleases, or uracil DNA deglycosylase (UDG), as shown in **FIG. 3**, upper part. This approach avoids sequence constraints of internal nicking recognition sites. The use of biotin-labeled primers can allow post-amplification return to single-stranded templates.

To avoid hairpins, synthesizing single-stranded transposons can be synthesized with each end forming a hairpin, as shown in **FIG. 3**, lower part. This hairpin prevents opposite-facing M-sites from forming hairpins. These hairpin structures are later

IDT Exhibit 1018.0327

processed to create active transposons with double-stranded M-sites and 5' phosphate and 3' OH groups.

— different first and second transposase recognition sequences

A particularly useful transposon sequence has different transposase recognition sequences (M and M') to reduce the subsequent formation of hairpin structures. For example, instead of having the same M-ends, a transposon can utilize an ME end (CTGTCTCTTATACACATCT or ME1) with an IE (inner end) end (CTGTCTCTTGATCAGATCT or ME2). An OE (outer ends) can also be combined with an IE. Preparation of the transposon is shown in **FIG. 4**. The degree of difference can be determined by screening for how much variation can be present in the ends to allow efficient transposition. A combination of two different ends that yield high transposition efficiency can be chosen from this screening.

The result is a transposome complex that has one transposon sequence and two transposases, as in **FIG. 5**. As shown, the transposase does not insert into single-stranded DNA such as the internal transposon sequence. After tagmentation, the single-stranded gap can be filled and ligated, as shown in **FIG. 6**. In this figure, there is no transposition into the transposon, and no ME-ME' hairpin is formed, since ME1 and ME2 are different.

— looped structures

A useful transposome is has one transposon that binds two transposase subunits forming a "looped" complex (**FIG. 7**). This configuration ensures that transposons are inserted into target DNA while maintaining ordering information of the original target DNA and without fragmenting the target DNA.

The transposon can be "pre-bent" in a configuration that favors the "looped" complex configuration, where transposon DNA ends have to come together and bend to derive the "looped" configuration. Without being tied to a particular mechanism, it is believed that the transposase subunits bind the ME-ends first, followed by dimerization of the subunits to form an active complex. It would be favorable for the two ends of the transposon to already be in a configuration that allows efficient intramolecular dimerization of the two bound transposase units, and can improve the yield of "looped" complex formation and stabilize the final complex. Such a configuration would also minimize undesired intermolecular dimerization products. To facilitate this bending

IDT Exhibit 1018.0328

process, the ends are positioned in close proximity, combining bending of transposon DNA with transposase complex formation.

Methods for bending DNA include surface binding, ion induction, chemical modification of DNA, protein-like bending with proteins such as Cro, histones, integration host factor (IHF), Fos, Yeast MCM1p, Jun, HU, dendrimer binding, and use of specific DNA sequences. For example, an IHF protein-DNA complex can be used to pre-bend because IHF recognizes and bends DNA at a specific sequence. Other proteins or other methods do not require a specific sequence. The introduction of one or more bends allows the creation of a specific 2- or 3-dimensional shape of the transposon that position the ME-ends in a desired position for "looped" complex formation. The agents responsible for bending can also prevent transposition into transposon-transposase complexes, and therefore favor insertion into target DNA.

Another approach favoring "looped" transposon-transposase complex formation is to relieve torsional stress, increasing rotational freedom in the transposon in order to allow the two ME-ends adapt a favorable orientation for "looped" complex formation. DNA molecules of certain sizes can be rigid ordinarily, so that nicks, gaps, single-stranded regions within the transposon can make the transposon flexible and adaptive. For example, formation of transposomes is enhanced when formed with transposons with a ~13bp region that is single-stranded, compared to its double-stranded version.

— pre-cleaved transposomes

In another alternative, pre-cleaved transposomes are prepared with a looped configuration (**FIG. 8**), ensuring that neighboring DNA fragments receive code combinations that can be unambiguously assembled at a later stage of the assay. As shown, the fragmentation maintains the code combination of N and N'. Subsequent fragmentation between the code-pair combination results in transposon-transposase complexes in which the transposon-ends are not linked to each other (*i.e.* a linear configuration). However, the complexes maintain the code-combinations as they were introduced in the looped configuration. Upon insertion of the fragmented transposon into the target, DNA fragments are created with codes on the ends, allowing code-mediated assembly of the sequenced DNA fragments/libraries.

The preparation of complexes with the looped configuration can be favored by various techniques. In one technique, the complexes are prepared in a dilute environment

IDT Exhibit 1018.0329

with excess transposase to favor formation of looped complexes. A subsequent concentration step prepares the complexes at a concentration desired for insertion. In another technique, the slow addition of transposons to transposases creates an environment with low amounts of transposon with excess transposase. This reaction condition ensures that each transposon will bind to two transposase units, forming the intra-molecular complex.

Nick-based methods

The methods herein are not limited to using the transposomes described above. Another embodiment can involve nicking of target DNA, which exploits the genomic DNA to provide the junction tags. Nicking methods can also be used in combination with transposon methods, such as preceding or following tagmentation. Any method that nicks a single strand of the target DNA may be used, preferably those that create random gaps in DNA. When random, the nicking methods can avoid any insertional bias that transposons may have. Alternatively, sequence-specific nicking enzymes can be used, or enzymes or chemical methods to remove a base or nucleotide from a single strand. Sequence-specific nicking enzymes include Nt.CviPII, Nb.BsmI, Nb.BbvCI, Nb.Bsr.DI, Nb.BtsI, Nt.BsmAI, Nt.BbvCI, Nt.BspQI, Nt.AlwI, and Nt.BstNBI (New England Biolabs).

As shown in **FIG. 9**, strand-displacement synthesis is used to copy identical flanking genomic DNA sequences (g-codes) at the end of formerly juxtaposed DNA fragments. The g-code is endogenous and therefore all the sequenced bases belong to the target DNA and not to any exogenously introduced code. After DNA fragments have obtained two identical g-codes, adapters or unique adapters (adapters with unique barcode) can be ligated to allow sequencing. The g-code sequences then serve as junction tags to assemble the individual fragments.

Alternatively, nicking followed by 3'-exo digestion can create gaps in genomic DNA without fragmenting the entire DNA sample, as shown in **FIG. 10**. The gaps can then be filled with code-containing oligonucleotides. For example, the oligonucleotides can be ligated to the two ends that were exposed by the nicking. Depending on the ligation method used (*e.g.* chemical or enzymatic), ligation may also be facilitated by phosphorylation of the oligo or the nicked end.

IDT Exhibit 1018.0330

Barcodes

Generally, a barcode can include one or more nucleotide sequences that can be used to identify one or more particular nucleic acids. The barcode can be an artificial sequence, or can be a naturally occurring sequence, such as a g-code, described herein. A barcode can comprise at least about 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20 or more consecutive nucleotides. In some embodiments, a barcode comprises at least about 10, 20, 30, 40, 50, 60, 70 80, 90, 100 or more consecutive nucleotides. In some embodiments, at least a portion of the barcodes in a population of nucleic acids comprising barcodes is different. In some embodiments, at least about 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90%, 95%, 99% of the barcodes are different. In more such embodiments, all of the barcodes are different. The diversity of different barcodes in a population of nucleic acids comprising barcodes can be randomly generated or non-randomly generated.

In some embodiments, a transposon sequence comprises at least one barcode. In some embodiments, a transposon sequence comprises a barcode comprising a first barcode sequence and a second barcode sequence. In some such embodiments, the first barcode sequence can be identified or designated to be paired with the second barcode sequence. For example, a known first barcode sequence can be known to be paired with a known second barcode sequence using a reference table comprising a plurality of first and second bar code sequences known to be paired to one another.

In another example, the first barcode sequence can comprise the same sequence as the second barcode sequence. In another example, the first barcode sequence can comprise the reverse complement of the second barcode sequence. In some embodiments, the first barcode sequence and the second barcode sequence are different ("bi-codes").

— bi-codes

As depicted in **FIG. 11**, transposons can be generated with either engineered bi-codes or random bi-codes. In the case of random bi-codes, the code pool may be down-selected (sub-sampled) to create unique code-code pairs. These unique bi-code combinations can then be identified by sequencing the transposons. The particular bi-code combinations (A_i and B_j) present in the final pool are determined and allow creation

IDT Exhibit 1018.0331

of a look-up table, which can be used to reassemble reads as described in the methods below.

An advantage of this approach is that not every code combination needs to be prepared individually. For example, *in situ* array technology can synthesize transposons or precursors with pre-designed code combinations in spatially separated locations, generating greater than 10^{18} different transposons or precursors in a single synthesis. To create unique bi-code maps, the codes are preferably down-selected so that a single first code is paired with only a single second code.

A method for reducing the presence of non-unique code-combinations (*e.g.*, Code1-Code2 and Code1-Code3 both present) is the use of "bottle-necking" or sub-sampling approaches. In these approaches, a large pool of many code combinations is generated. For example, two 15bp random sequences (4^{15}) can result in more than 10^{18} different transposons. Taking a fraction of this high-complexity pool results in unique code-combinations for the sub-sampled fraction.

In another embodiment, a transposon sequence comprising a double-strand nucleic acid, each strand can comprise a different barcode sequence. As will be understood, such a transposon sequence can be useful to inhibit insertion of other sequences into the transposon sequence by transposition. Such transposon sequences can also be used to tag particular strands of a double-stranded target nucleic acid, for example in methods that include haplotypes sequencing. An exemplary transposon sequence comprising a plurality of barcode sequences is depicted in **FIG. 2.**

— half-transposons

In another embodiment, code combinations are introduced after formation of the transposon-transposase complex. Complexes with non-linked ME regions can generate linked code information after transposon insertion and subsequent downstream assay steps. Complexes can be prepared as illustrated in **FIG. 12**, in which transposon ends are not linked. For example, complexes can be formed from two "half-transposons", each containing one ME region and a random or specific code region. This generates transposase complexes with randomly inserted transposons ends; no prior knowledge of pairing between the two half-codes is required. Each transposon contains a ME sequence for transposase recognition, a random or specific code, and an endonuclease fragmentation system. Upon insertion, the transposase enzyme can prevent target DNA

IDT Exhibit 1018.0332

fragmentation by binding both transposon ends. The ends of transposons can now be ligated to each other via the sticky ends. Upon nicking each strand (*e.g.,* by a sequence-specific fragmentation system, such as a restriction endonuclease), sticky code-tagged ends can be created that link one DNA fragment to the other. The codes generated at each junction are reverse complements to one another. Denaturation creates DNA fragments in which each DNA fragment contains a linked code with the neighboring fragment. An advantage of this embodiment is that transposon-transposase complexes with specific or random codes can readily be prepared, without the need for making complexes with looped configurations. Furthermore in this embodiment, the code combination does not have to be introduced at the transposase complex level, but can be generated at a later stage in the assay.

As will be described further herein, paired first and second barcode sequences can be used to identify different nucleic acids comprising barcodes linked with one another.

Linkers

Some embodiments include transposon sequences comprising a first barcode sequence and a second barcode sequence having a linker disposed therebetween. In other embodiments, the linker can be absent, or can be the sugar-phosphate backbone that connects one nucleotide to another. The linker can comprise, for example, one or more of a nucleotide, a nucleic acid, a non-nucleotide chemical moiety, a nucleotide analogue, amino acid, peptide, polypeptide, or protein. In preferred embodiments, a linker comprises a nucleic acid. The linker can comprise at least about 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, or more nucleotides. In some embodiments, a linker can comprise at least about 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, or more nucleotides.

— nonamplifiable linkers

In another embodiment, non-amplifiable linkers can be introduced between ME regions (**FIG. 13**), which can be useful in methods using bi-codes, as in **FIG. 14.** Prior to Round 1 in the figure, it is preferable to perform an amplification with P1 primers to increase the number of copies, and to ensure that all templates in the population are sequenceable. Examples of non-amplifiable linkers include organic chemistry linkers such as alkyl, propyl, PEG; non-natural bases such as IsoC, isoG; or any group that does

IDT Exhibit 1018.0333

not amplify in DNA-based amplification schemes. For example, transposons containing isoC, isoG pairs can be amplified with dNTPs mixtures lacking a complementary isoG and isoC, ensuring that no amplification occurs across the inserted transposons.

In this method, transposons are inserted into target DNA with the following configuration: ME-Code*N*-P1-linker-P1'-Code*N*'-ME so that code*N* and Code*N*' generate the code combination together. P1 is a primer and P1' is its complement. PCR amplification with P1 or P1' (*e.g.,* generates DNA fragments that are code-tagged. Fragmentation at the linker site or a non-amplifiable linker can reduce undesirable amplification across transposons, which may reduce the yield of the libraries or generate undesired products.

Fragmentation sites

In some embodiments, the linker can comprise a fragmentation site. A fragmentation site can be used to cleave the physical, but not the informational association between a first barcode sequence and a second barcode sequence. Cleavage may be by biochemical, chemical or other means. In some embodiments, a fragmentation site can include a nucleotide or nucleotide sequence that may be fragmented by various means. For example, a fragmentation site may be a substrate for an enzyme, such as a nuclease, that will cleave the physical association between a first barcode sequence and a second barcode sequence. For example, the fragmentation site comprises a restriction endonuclease site and may be cleaved with an appropriate restriction endonuclease. In another example, a fragmentation site can comprise at least one ribonucleotide in a nucleic acid that may otherwise comprise deoxyribonucleotides and may be cleaved with an RNAse. Chemical cleavage agents capable of selectively cleaving the phosphodiester bond between a deoxyribonucleotide and a ribonucleotide include metal ions, for example rare-earth metal ions (*e.g.,* La^{3+}, particularly Tm^{3+}, Yb^{3+} or Lu^{3+} (Chen *et al.* Biotechniques. 2002, 32: 518-520; Komiyama *et al.* Chem. Commun. 1999, 1443-1451)), Fe(3) or Cu(3), or exposure to elevated pH, *e.g.,* treatment with a base such as sodium hydroxide. As used herein, selective cleavage of the phosphodiester bond between a deoxyribonucleotide and a ribonucleotide can refer to the chemical cleavage agent is not capable of cleaving the phosphodiester bond between two deoxyribonucleotides under the same conditions.

IDT Exhibit 1018.0334

In another example, the fragmentation site can comprise one or more recognition sequences for a nickase, that is, a nicking endonuclease that breaks one strand of a double-stranded nucleic acid. Thus, the fragmentation site can comprise a first nickase recognition sequence, a second nickase recognition sequence. The cut site for each recognition sequence can be the same site or different site.

In another example, a fragmentation site can include one or more nucleotide analogues that comprise an abasic site and permits cleavage at the fragmentation site in the presence of certain chemical agents, such as polyamine, N,N′-dimethylethylenediamine (DMED) (U.S. Patent Publication No. 2010/0022403). In some embodiments, an abasic site may be created within a fragmentation site by first providing a fragmentation site comprising a deoxyuridine (U) of a double stranded nucleic acid. The enzyme uracil DNA glycosylase (UDG) may then be used to remove the uracil base, generating an abasic site on one strand. The polynucleotide strand including the abasic site may then be cleaved at the abasic site by treatment with endonuclease (e.g. Endo IV endonuclease, AP lyase, FPG glycosylase/AP lyase, Endo VIII glycosylase/AP lyase), heat or alkali. Abasic sites may also be generated at nucleotide analogues other than deoxyuridine and cleaved in an analogous manner by treatment with endonuclease, heat or alkali. For example, 8-oxo-guanine can be converted to an abasic site by exposure to FPG glycosylase. Deoxyinosine can be converted to an abasic site by exposure to AlkA glycosylase. The abasic sites thus generated may then be cleaved, typically by treatment with a suitable endonuclease (e.g. Endo IV, AP lyase). (U.S. Patent Publication No. 2011/0014657).

In another example, a fragmentation site may include a diol linkage which permits cleavage by treatment with periodate (*e.g.,* sodium periodate). In another example, a fragmentation site may include a disulphide group which permits cleavage with a chemical reducing agent, e.g. Tris (2-carboxyethyl)-phosphate hydrochloride (TCEP).

In some embodiments, a fragmentation site may include a cleavable moiety that may be subject to photochemical cleavage. Photochemical cleavage encompasses any method which utilizes light energy in order to achieve cleavage of nucleic acids, for example, one or both strands of a double-stranded nucleic acid molecule. A site for photochemical cleavage can be provided by a non-nucleotide chemical moiety in a nucleic acid, such as phosphoramidite [4-(4,4'-dimethoxytrityloxy)butyramidomethyl)-1-

IDT Exhibit 1018.0335

(2-nitrophenyl)-ethyl]-2-cyanoethyl-(N,N-diisopropyl)-phosphoramidite) (Glen Research, Sterling, Va., USA, Cat No. 10-4913-XX).

In some embodiments, a fragmentation site can include a peptide, for example, conjugate structure in which a peptide molecule is linked to a nucleic acid. The peptide molecule can subsequently be cleaved by a peptidase enzyme of the appropriate specificity, or any other suitable means of non-enzymatic chemical or photochemical cleavage. In some embodiments, a conjugate between peptide and nucleic acid will be formed by covalently linking a peptide to a nucleic acid, *e.g.*, a strand of a double-stranded nucleic acid. Conjugates between a peptide and nucleic acid can be prepared using techniques generally known in the art. In one such technique the peptide and nucleic acid components of the desired amino acid and nucleotide sequence can be synthesized separately, e.g. by standard automated chemical synthesis techniques, and then conjugated in aqueous/organic solution. By way of example, the OPeC™ system commercially available from Glen Research is based on the native ligation of an N-terminal thioester-functionalized peptide to a 5′-cysteinyl oligonucleotide.

Primer sites

In some embodiments, a linker can be a "sequencing adaptor" or "sequencing adaptor site", that is to say a region that comprises one or more sites that can hybridize to a primer. In some embodiments, a linker comprises at least a first primer site. In some embodiments, a linker comprises at least a first primer site and a second primer site. The orientation of the primer sites in such embodiments can be such that a primer hybridizing to the first primer site and a primer hybridizing to the second primer site are in the same orientation, or in different orientations. In one embodiment, the primer sequence in the linker can be complementary to a primer used for amplification. In another embodiment, the primer sequence is complementary to a primer used for sequencing.

In some embodiments, a linker can include a first primer site, a second primer site having a non-amplifiable site disposed therebetween. The non-amplifiable site is useful to block extension of a polynucleotide strand between the first and second primer sites, wherein the polynucleotide strand hybridizes to one of the primer sites. The non-amplifiable site can also be useful to prevent concatamers. Examples of non-amplifiable sites include a nucleotide analogue, non-nucleotide chemical moiety, amino-acid, peptide,

IDT Exhibit 1018.0336

and polypeptide. In some embodiments, a non-amplifiable site comprises a nucleotide analogue that does not significantly base-pair with A, C, G or T.

Some embodiments include a linker comprising a first primer site, a second primer site having a fragmentation site disposed therebetween.

Other embodiments can use a forked or Y-shaped adapter design useful for directional sequencing, as described in U.S. Patent No. 7,741,463. An example is shown in **FIG. 15.**

Affinity tags

In some embodiments, a linker can comprise an affinity tag. Affinity tags can be useful for the bulk separation of target nucleic acids hybridized to hybridization tags. As used herein, the term "affinity tag" and grammatical equivalents can refer to a component of a multi-component complex, wherein the components of the multi-component complex specifically interact with or bind to each other. For example an affinity tag can include biotin or His that can bind streptavidin or nickel, respectively. Other examples of multiple-component affinity tag complexes include, ligands and their receptors, for example, avidin-biotin, streptavidin-biotin, and derivatives of biotin, streptavidin, or avidin, including, but not limited to, 2-iminobiotin, desthiobiotin, NeutrAvidin (Molecular Probes, Eugene, Oreg.), CaptAvidin (Molecular Probes), and the like; binding proteins/peptides, including maltose-maltose binding protein (MBP), calcium-calcium binding protein/peptide (CBP); antigen-antibody, including epitope tags, including c-MYC (*e.g.,* EQKLISEEDL (SEQ ID NO: 3)), HA (*e.g.,* YPYDVPDYA (SEQ ID NO: 4)), VSV-G (*e.g.,* YTDIEMNRLGK (SEQ ID NO: 5)), HSV (*e.g.,* QPELAPEDPED (SEQ ID NO: 6)), V5 (*e.g.,* GKPIPNPLLGLDST (SEQ ID NO: 7)), and FLAG Tag™. (*e.g.,* DYKDDDDKG (SEQ ID NO: 8)), and their corresponding anti-epitope antibodies; haptens, for example, dinitrophenyl and digoxigenin, and their corresponding antibodies; aptamers and their corresponding targets; poly-His tags (*e.g.,* penta-His and hexa-His) and their binding partners including corresponding immobilized metal ion affinity chromatography (IMAC) materials and anti-poly-His antibodies; fluorophores and anti-fluorophore antibodies; and the like.

IDT Exhibit 1018.0337

Reporter moieties

In some embodiments, a linker can comprise a reporter moiety. As used herein, the term "reporter moiety" and grammatical equivalents can refer to any identifiable tag, label, or group. The skilled artisan will appreciate that many different species of reporter moieties can be used with the methods and compositions described herein, either individually or in combination with one or more different reporter moieties. In certain embodiments, a reporter moiety can emit a signal. Examples of signals fluorescent, a chemiluminescent, a bioluminescent, a phosphorescent, a radioactive, a calorimetric, or an electrochemiluminescent signals. Example reporter moieties include fluorophores, radioisotopes, chromogens, enzymes, antigens including epitope tags, semiconductor nanocrystals such as quantum dots, heavy metals, dyes, phosphorescence groups, chemiluminescent groups, electrochemical detection moieties, binding proteins, phosphors, rare earth chelates, transition metal chelates, near-infrared dyes, electrochemiluminescence labels, and mass spectrometer compatible reporter moieties, such as mass tags, charge tags, and isotopes. More reporter moieties that may be used with the methods and compositions described herein include spectral labels such as fluorescent dyes (*e.g.,* fluorescein isothiocyanate, Texas red, rhodamine, and the like), radiolabels (*e.g.,* ^{3}H, ^{125}I, ^{35}S, ^{14}C, ^{32}P, ^{33}P, etc.), enzymes (*e.g.,* horseradish peroxidase, alkaline phosphatase etc.) spectral calorimetric labels such as colloidal gold or colored glass or plastic (e.g. polystyrene, polypropylene, latex, etc.) beads; magnetic, electrical, thermal labels; and mass tags. Reporter moieties can also include enzymes (horseradish peroxidase, etc.) and magnetic particles. More reporter moieties include chromophores, phosphors and fluorescent moieties, for example, Texas red, dixogenin, biotin, 1- and 2-aminonaphthalene, p,p'-diaminostilbenes, pyrenes, quaternary phenanthridine salts, 9-aminoacridines, p,p'-diaminobenzophenone imines, anthracenes, oxacarbocyanine, merocyanine, 3-aminoequilenin, perylene, bis-benzoxazole, bis-p-oxazolyl benzene, 1,2-benzophenazin, retinol, bis-3-aminopyridinium salts, hellebrigenin, tetracycline, sterophenol, benzimidazolylphenylamine, 2-oxo-3-chromen, indole, xanthen, 7-hydroxycoumarin, phenoxazine, calicylate, strophanthidin, porphyrins, triarylmethanes and flavin. Individual fluorescent compounds which have functionalities for linking to an element desirably detected in an apparatus or assay provided herein, or which can be modified to incorporate such functionalities include, *e.g.,* dansyl chloride; fluoresceins such as 3,6-dihydroxy-9-phenylxanthydrol; rhodamineisothiocyanate; N-phenyl 1-amino-

IDT Exhibit 1018.0338

8-sulfonatonaphthalene; N-phenyl 2-amino-6-sulfonatonaphthalene; 4-acetamido-4-isothiocyanato-stilbene-2,2'-disulfonic acid; pyrene-3-sulfonic acid; 2-toluidinonaphthalene-6-sulfonate; N-phenyl-N-methyl-2-aminoaphthalene-6-sulfonate; ethidium bromide; stebrine; auromine-0,2-(9'-anthroyl)palmitate; dansyl phosphatidylethanolamine; N,N'-dioctadecyl oxacarbocyanine: N,N'-dihexyl oxacarbocyanine; merocyanine, 4-(3'-pyrenyl)stearate; d-3-aminodesoxy-equilenin; 12-(9'-anthroyl)stearate; 2-methylanthracene; 9-vinylanthracene; 2,2'(vinylene-p-phenylene)bisbenzoxazole; p-bis(2- -methyl-5-phenyl-oxazolyl))benzene; 6-dimethylamino-1,2-benzophenazin; retinol; bis(3'-aminopyridinium) 1,10-decandiyl diiodide; sulfonaphthylhydrazone of hellibrienin; chlorotetracycline; N-(7-dimethylamino-4-methyl-2-oxo-3-chromenyl)maleimide; N-(p-(2benzimidazolyl)-phenyl)maleimide; N-(4-fluoranthyl)maleimide; bis(homovanillic acid); resazarin; 4-chloro7-nitro-2,1,3-benzooxadiazole; merocyanine 540; resorufin; rose bengal; 2,4-diphenyl-3(2H)-furanone, fluorescent lanthanide complexes, including those of Europium and Terbium, fluorescein, rhodamine, tetramethylrhodamine, eosin, erythrosin, coumarin, methyl-coumarins, quantum dots (also referred to as "nanocrystals": see U.S. Patent No. 6,544,732), pyrene, Malachite green, stilbene, Lucifer Yellow, Cascade Blue™, Texas Red, Cy dyes (Cy3, Cy5, etc.), Alexa Fluor® dyes, phycoerythin, bodipy, and others described in the 6th Edition of the *Molecular Probes Handbook* by Richard P. Haugland,.

Certain methods of making transposon sequences

The transposon sequences provided herein can be prepared by a variety of methods. Example methods include direct synthesis and hairpin extension methods. In some embodiments, transposon sequences may be prepared by direct synthesis. For example, a transposon sequence comprising a nucleic acid may be prepared by methods comprising chemical synthesis. Such methods are well known in the art, *e.g.,* solid phase synthesis using phophoramidite precursors such as those derived from protected 2′-deoxynucleosides, ribonucleosides, or nucleoside analogues.

In some embodiments, a transposon sequence can be prepared by hairpin extension. In some such embodiments, a portion of a transposon sequence may be prepared by chemical synthesis and extended by hairpin extension. In an example embodiment, a precursor transposon sequence comprising a polynucleotide may include a first transposase recognition site, and a first barcode sequence. The precursor transposon

IDT Exhibit 1018.0339

sequence may be extended using an appropriate nucleic acid polymerase by hairpin extension, thereby preparing a hairpin structure comprising a first transposase recognition site, a first barcode sequence, a second barcode sequence, and a second transposase recognition site. In such a transposon sequence, the first transposase recognition site and first barcode sequence can have the reverse complement sequence of, and the second transposase recognition site and the second barcode sequence, respectively.

Some methods of preparing transposons sequences can include preparing barcode sequences. Barcode sequences can be generated randomly and non-randomly. Some barcode sequences may or may not include sequences likely to be found in a target nucleic acid. Some barcode sequences may or may not include restriction sites. In some embodiments, at least a portion of a barcode sequence can be generated randomly. In some embodiments a barcode sequence can be generated using combinatorial methods. In some such methods, barcode sequences can comprise one or more subunits comprising one or more consecutive nucleotides. As will be understood, in some embodiments, a subunit can comprise a nucleotide analogue and/or a nucleotide comprising a label. At least a portion of a barcode sequence can comprise at least one subunit. Barcode sequences comprising at least one subunit can be generated randomly or non-randomly. In some embodiments, at least a portion of a barcode sequences comprising at least one subunit can be generated randomly or non-randomly. As will be understood, in some embodiments, at least a portion of a barcode sequence is known.

Primer sites may be ligated to the ends of the hairpin structure in order to generate a complementary strand to the single-strand of the hairpin structure. An exemplary embodiment of a method of making a transposon sequence is depicted in **FIG. 16**. It will be understood that the primer sites introduced to generate the complementary strand may be removed by a transposase during a transposition reaction.

In some embodiments, methods of making transposon sequences can include incorporating sequences using amplification methods. Templates for amplification that include barcode sequences can be prepared by a variety of systems, for example, using *in situ* oligonucleotide arrays. Some methods of making transposon sequences include the use of tailed-oligonucleotides to incorporate sequences into a transposon sequence. An exemplary embodiment is depicted in **FIG. 17,** where tailed-oligonucleotides comprising mosaic elements (ME) are used to prepare a transposon sequence comprising a barcode comprising Code X and Code Y, and a linker comprising a fragmentation site.

IDT Exhibit 1018.0340

Target nucleic acids

A target nucleic acid can include any nucleic acid of interest. Target nucleic acids can include DNA, RNA, peptide nucleic acid, morpholino nucleic acid, locked nucleic acid, glycol nucleic acid, threose nucleic acid, mixtures thereof, and hybrids thereof. In a preferred embodiment, genomic DNA fragments or amplified copies thereof are used as the target nucleic acid. In another preferred embodiment, mitochondrial or chloroplast DNA is used.

A target nucleic acid can comprise any nucleotide sequence. In some embodiments, the target nucleic acid comprises homopolymer sequences. A target nucleic acid can also include repeat sequences. Repeat sequences can be any of a variety of lengths including, for example, 2, 5, 10, 20, 30, 40, 50, 100, 250, 500, 1000 nucleotides or more. Repeat sequences can be repeated, either contiguously or non-contiguously, any of a variety of times including, for example, 2, 3, 4, 5, 6, 7, 8, 9, 10, 15, 20 times or more.

Some embodiments described herein can utilize a single target nucleic acid. Other embodiments can utilize a plurality of target nucleic acids. In such embodiments, a plurality of target nucleic acids can include a plurality of the same target nucleic acids, a plurality of different target nucleic acids where some target nucleic acids are the same, or a plurality of target nucleic acids where all target nucleic acids are different. Embodiments that utilize a plurality of target nucleic acids can be carried out in multiplex formats so that reagents are delivered simultaneously to the target nucleic acids, for example, in one or more chambers or on an array surface. In some embodiments, the plurality of target nucleic acids can include substantially all of a particular organism's genome. The plurality of target nucleic acids can include at least a portion of a particular organism's genome including, for example, at least about 1%, 5%, 10%, 25%, 50%, 75%, 80%, 85%, 90%, 95%, or 99% of the genome. In particular embodiments the portion can have an upper limit that is at most about 1%, 5%, 10%, 25%, 50%, 75%, 80%, 85%, 90%, 95%, or 99% of the genome

Target nucleic acids can be obtained from any source. For example, target nucleic acids may be prepared from nucleic acid molecules obtained from a single organism or from populations of nucleic acid molecules obtained from natural sources that include one or more organisms. Sources of nucleic acid molecules include, but are not limited to, organelles, cells, tissues, organs, or organisms. Cells that may be used as sources of target

IDT Exhibit 1018.0341

nucleic acid molecules may be prokaryotic (bacterial cells, for example, *Escherichia, Bacillus, Serratia, Salmonella, Staphylococcus, Streptococcus, Clostridium, Chlamydia, Neisseria, Treponema, Mycoplasma, Borrelia, Legionella, Pseudomonas, Mycobacterium, Helicobacter, Erwinia, Agrobacterium, Rhizobium,* and *Streptomyces* genera); archeaon, such as crenarchaeota, nanoarchaeota or euryarchaeotia; or eukaryotic such as fungi, (for example, yeasts), plants, protozoans and other parasites, and animals (including insects (for example, *Drosophila* spp.), nematodes (*e.g., Caenorhabditis elegans*), and mammals (for example, rat, mouse, monkey, non-human primate and human)).

Methods of preparing template nucleic acids

Some embodiments include methods of preparing template nucleic acids. As used herein, the term "template nucleic acid" can refer to a target nucleic acid, a fragment thereof, or any copy thereof comprising at least one transposon sequence, a fragment thereof, or any copy thereof. Accordingly, some methods of preparing template nucleic acids include inserting a transposon sequence into a target nucleic acid, thereby preparing a template nucleic acid. Some methods of insertion include contacting a transposon sequence provided herein with a target nucleic acid in the presence of an enzyme, such as a transposase or integrase, under conditions sufficient for the integration of the transposon sequence into the target nucleic acid.

In some embodiments, insertion of transposon sequences into a target nucleic acid can be non-random. In some embodiments, transposon sequences can be contacted with target nucleic acids comprising proteins that inhibit integration at certain sites. For example, transposon sequences can be inhibited from integrating into genomic DNA comprising proteins, genomic DNA comprising chromatin, genomic DNA comprising nucleosomes, or genomic DNA comprising histones. In some embodiments, transposon sequences can be associated with affinity tags in order to integrate the transposon sequence at a particular sequence in a target nucleic acid. For example, a transposon sequence may be associated with a protein that targets specific nucleic acid sequences, *e.g.,* histones, chromatin-binding proteins, transcription factors, initiation factors, etc., and antibodies or antibody fragments that bind to particular sequence-specific nucleic-acid-binding proteins. In an exemplary embodiment, a transposon sequence is associated with an affinity tag, such as biotin; the affinity tag can be associated with a nucleic-acid-binding protein.

IDT Exhibit 1018.0342

It will be understood that during integration of some transposon sequences into a target nucleic acid, several consecutive nucleotides of the target nucleic acid at the integration site are duplicated in the integrated product. Thus the integrated product can include a duplicated sequence at each end of the integrated sequence in the target nucleic acid. An example of such a duplication event is depicted in **FIG. 18**. As used herein, the term "host tag" can refer to a target nucleic acid sequence that is duplicated at each end of an integrated transposon sequence. Single-stranded portions of nucleic acids that may be generated by the insertion of transposon sequences can be repaired by a variety of methods well known in the art, for example by using ligases, oligonucleotides and/or polymerases.

In some embodiments, a plurality of the transposon sequences provided herein is inserted into a target nucleic acid. Some embodiments include selecting conditions sufficient to achieve integration of a plurality of transposon sequences into a target nucleic acid such that the average distance between each integrated transposon sequence comprises a certain number of consecutive nucleotides in the target nucleic acid.

In some embodiments, conditions for insertion of transposon sequences are sufficient to reduce the likelihood of forming concatameric complexes comprising a transposase associated with more than one transposon sequence. In one example, complexes comprising a transposase and a transposon sequence can be formed under dilute conditions; subsequent steps of inserting the transposon sequences into a target nucleic acid may be carried out at higher concentrations of transposase/transposon sequence complex. In another example, transposase/transposon sequence complexes can be prepared by contacting a circular transposon sequence with a transposase. As will be understood, the transposon sequence may be linearized during formation of the transposase/transposon sequence complex. In another example, a transposase/transposon sequence complex may be prepared by preparing complexes comprising partial transposon sequences comprising a transposon recognition site, and contacting the partial sequences with transposase monomers. Two partial transposon sequences may be ligated to one another to prepare a whole transposon sequence associated with a transposase comprising a dimer.

Some embodiments include selecting conditions sufficient to achieve insertion of a transposon sequence into a target nucleic acid, but not into another transposon sequence. A variety of methods can be used to reduce the likelihood that a transposon

IDT Exhibit 1018.0343

sequence inserts into another transposon sequence. For example, transposon sequences can comprise thiophosphate-modified nucleic acids. In another example, a transposon sequence can comprise a DNA/RNA hybrid, such as an RNA transposon sequence comprising DNA transposase recognition sites. In another example, a transposon sequence comprises a single-stranded sequence, further comprising double-stranded transposase recognition sites. It will be appreciated that more methods can include transposon sequences comprising single-stranded nucleic acids to inhibit insertion into the transposon sequence; transposon sequences comprising RNA to inhibit insertion into the transposon sequence; and transposon sequences associated with nucleic acid binding proteins to inhibit insertion into the transposon sequence.

In some embodiments, conditions may be selected so that the average distance in a target nucleic acid between integrated transposon sequences is at least about 5, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, or more consecutive nucleotides. In some embodiments, the average distance in a target nucleic acid between integrated transposon sequences is at least about 100, 200, 300, 400, 500, 600, 700, 800, 900, 1000, or more consecutive nucleotides. In some embodiments, the average distance in a target nucleic acid between integrated transposon sequences is at least about 1 kb, 2 kb, 3 kb, 4 kb, 5 kb, 6 kb, 7 kb, 8 kb, 90 kb, 100 kb, or more consecutive nucleotides. In some embodiments, the average distance in a target nucleic acid between integrated transposon sequences is at least about 100 kb, 200 kb, 300 kb, 400 kb, 500 kb, 600 kb, 700 kb, 800 kb, 900 kb, 1000 kb, or more consecutive nucleotides. As will be understood, some conditions that may be selected include contacting a target nucleic acid with a certain number of transposon sequences.

Some embodiments include selecting conditions sufficient to achieve at least a portion of transposon sequences integrated into a target nucleic acid are different. In preferred embodiments, each transposon sequence integrated into a target nucleic acid is different. Some conditions that may be selected to achieve a certain portion of transposon sequences integrated into a target sequences that are different include selecting the degree of diversity of the population of transposon sequences. As will be understood, the diversity of transposon sequences arises in part due to the diversity of the barcodes of such transposon sequences. Accordingly, some embodiments include providing a population of transposon sequences in which at least a portion of the barcodes are different. In some embodiments, at least about 10%, 20%, 30%, 40%, 50%, 60%, 70%,

IDT Exhibit 1018.0344

80%, 90%, 95%, 98%, 99%, or 100% of barcodes in a population of transposon sequences are different.

Some embodiments of preparing a template nucleic acid can include copying the sequences comprising the target nucleic acid. For example, some embodiments include hybridizing a primer to a primer site of a transposon sequence integrated into the target nucleic acid. In some such embodiments, the primer can be hybridized to the primer site and extended. The copied sequences can include at least one barcode sequence and at least a portion of the target nucleic acid. In some embodiments, the copied sequences can include a first barcode sequence, a second barcode sequence, and at least a portion of a target nucleic acid disposed therebetween. In some embodiments, at least one copied nucleic acid can include at least a first barcode sequence of a first copied nucleic acid that can be identified or designated to be paired with a second barcode sequence of a second copied nucleic acid. In some embodiments, the primer can include a sequencing primer. In some embodiments sequencing data is obtained using the sequencing primer.

Some embodiments of preparing a template nucleic acid can include amplifying sequences comprising at least a portion of one or more transposon sequences and at least a portion of a target nucleic acid. In some embodiments, at least a portion of a target nucleic acid can be amplified using primers that hybridize to primer sites of integrated transposon sequences integrated into a target nucleic acid. In some such embodiments, an amplified nucleic acid can include a first barcode sequence, and second barcode sequence having at least a portion of the target nucleic acid disposed therebetween. In some embodiments, at least one amplified nucleic acid can include at least a first barcode sequence of a first amplified nucleic acid that can be identified to be paired with a second barcode sequence of a second amplified sequence.

— symmetrical transposomes, adding sequences by amplification

In an embodiment, transposomes are used that have symmetrical transposable end sequences, as exemplified in **FIG. 19**. Each tagmented fragment therefore contains identical ends, lacking directionality. A single primer PCR, using the transposon end sequences, can then be employed to amplify the template copy number from $2n$ to $2n*2^x$ where x corresponds to the number of PCR cycles. In a subsequent step, PCR with primers can add additional sequences, such as sequencing adapter sequences.

IDT Exhibit 1018.0345

Some embodiments of preparing a template nucleic acid can include fragmenting a target nucleic acid comprising transposon sequences. Methods of fragmenting nucleic acids are well known in the art. In some embodiments, a nucleic acid comprising transposon sequences can be fragmented at random positions along the length of the nucleic acid. In some embodiments, a target nucleic acid comprising transposon sequences can be fragmented at the fragmentation sites of the transposon sequences. In some embodiments, insertion of a transposon sequence can include the duplication of the insertion site so that the inserted transposon sequence is disposed between duplicated single-stranded sequences (*See, e.g.,* **FIG. 18**). In some embodiments, a polymerase may be used to cleave the fragmentation site. Examples of such polymerases include strand-displacing nucleic acid polymerases.

Further embodiments of preparing a template nucleic acid that include fragmenting a target nucleic acid comprising transposon sequences can also include amplifying the fragmented nucleic acids. In some embodiments, the fragmented nucleic acids can be amplified using primers that hybridize to primer sites of transposon sequences. In more embodiments, primer sites can be ligated to the ends of the fragmented nucleic acids. In some such embodiments, the fragmented nucleic acids with ligated primer sites can be amplified from such primer sites.

Some embodiments include reducing the complexity of a library of template nucleic acids. A complexity-reduction step can be performed before or after the fragmentation step in the method. For example, the target nucleic acid comprising the transposon sequences can be diluted so that a small number or a single molecule represents the target diluted before performing subsequent steps.

FIG. 20 depicts an exemplary embodiment of a contacting a target nucleic acid with a library of transposons, which integrates into the target nucleic acid. **FIG. 21** depicts an optional step of fragmenting a linker into two parts, or an optional step of virtually fragmenting the linker by amplification using a primer A′ and/or a primer B′. Fragmentation of the linker generates a library of template nucleic acids. In this figure, the filled-in region is indicated by a thick line.

In some embodiments, it can be advantageous for each template nucleic acid to incorporate at least one universal primer site. For example, a template nucleic acid can include first end sequences that comprise a first universal primer site, and second end sequences that comprise a second universal primer site. Universal primer sites can have

IDT Exhibit 1018.0346

various applications, such as amplifying, sequencing, and/or identifying one or more template nucleic acids. The first and second universal primer sites can be the same, substantially similar, similar, or different. In some embodiments, in order to prepare a template nucleic acid comprising a first universal primer site and a second universal primer site, a transposon sequence is prepared that includes a first transposase recognition site, a second transposase recognition site, a barcode disposed therebetween, wherein the barcode comprises a first barcode sequence and a second barcode sequence, separated by a linker. The linker includes a first primer site and a second primer site with a fragmentation site therebetween. The first primer site can comprise sequences that are the reverse complement of sequences within the second primer. In some embodiments, the first primer site comprises sequences having dyad symmetry to sequences within the second primer. In some embodiments, the first primer site comprises sequences having C_2 symmetry to sequences within the second primer. A plurality of transposons may be inserted into a target nucleic acid by transposition in the presence of a transposase. The incorporated sequences may be cleaved to yield a plurality of target nucleic acids, each comprising the first primer site and second primer site. In some embodiments, a first universal primer site and a second universal primer site can be incorporated into each template nucleic acid by a variety of methods. For example, template nucleic acids can be amplified using the first primer site and second primer site using tailed-oligonucleotides. As is understood in the art, a tailed-oligonucleotide can include sequences complementary to a primer site and additional sequences. In an example embodiment, a first tailed-oligonucleotide comprises sequences complementary to a first primer site and sequences for a first universal primer site, and a second tailed-oligonucleotide comprises sequences complementary to a second primer site and sequences for a second universal primer site. **FIG. 22** depicts an example embodiment series of rounds of extending template using tailed-oligonucleotides to obtain a population of template nucleic acids comprising template nucleic acids that include a first universal primer site and a second universal primer site. With respect to, in step 1, oligonucleotides comprising either a first universal primer sequence (A) or a second universal primer sequence (B) anneal to a primer site (P) of a template nucleic acid, and are extended. In step 2, the extension products provide templates for a further extension step. Alternatively, the extension products can be amplified with the use of additional oligonucleotides. The products of step 2 are depicted in step 3 and include nucleic acid

IDT Exhibit 1018.0347

sequences comprising first universal primer sites only, second universal primer sites only, or first and second universal primer sites. As will be understood, nucleic acid sequences comprising a first universal primer site and second universal primer site may be used in further sequencing methods.

It will be understood that in some embodiments, the vast number of available barcodes permits each template nucleic acid molecule to comprise a unique identification. Unique identification of each molecule in a mixture of template nucleic acids can be used in several applications to identify individual nucleic acid molecules, in samples having multiple chromosomes, genomes, cells, cell types, cell disease states, and species, for example in haplotype sequencing, parental allele discrimination, metagenomic sequencing, and sample sequencing of a genome.

Tagging molecules individually

The present invention also provides methods for tagging molecules uniquely so individual molecules can be tracked and identified, so that bulk data can be deconvoluted back to the individual molecule. The ability to distinguish individual molecules and relate the information back to the molecule of origin is especially important when processes from original molecule to final product change the (stoichiometric) representation of the original population. For example, amplification leads to duplication (*e.g.,* PCR duplicates or biased amplification) that can skew the original representation. This can alter the methylation state call, copy number, allelic ratio due to non-uniform amplification and/or amplification bias. By identifying individual molecules, code-tagging distinguishes between identical molecules after processing. As such, duplications, and amplification bias can be filtered out, allowing accurate determination of the original representation of a molecule or population of molecules.

An advantage to this method is molecules that are otherwise identical in the original pool become uniquely identified by virtue of their tagging. In further downstream analysis these molecules can now be distinguished. This technique can be exploited in assay schemes in which amplification is employed. For example, amplification is known to distort the original representation of a mixed population of molecules. If unique tagging were not employed, the original representation (such as copy number or allelic ratio) would need to account for the biases (known or unknown) for each molecule in the representation. With unique tagging, the representation can

IDT Exhibit 1018.0348

accurately be determined by removing duplicates and counting the original representation of molecules, each having a unique tag. Thus, the cDNAs can be amplified and sequenced, without fear of bias, followed by filtering the data so that only authentic sequences or sequences of interest are selected for further analysis.

It is preferred to tag the original population in the early stages of the assay, although tagging can occur at later stages if the earlier steps do not introduce bias or are not important. In any of these applications, the complexity of the barcode sequences should be larger than the number of individual molecules to be tagged. This ensures that different target molecules receive different and unique tags. As such, a pool of random oligonucleotides of a certain length (5, 10, 20, 30, 40, 50, 100, 200) is desirable. Random pool of tags represents a large complexity of tags with code space 4^n where n is the number of nucleotides. Additional codes (whether designed or random) can be incorporated at different stages to serve as a further check, such as a parity check for error correction.

In one embodiment, individual molecules (such as target DNA) are attached to unique labels, such as unique oligo sequences, barcodes, as shown in **FIG. 23**. Attachment can occur through ligation, coupling chemistry, adsorption, etc. Other means include amplification (such as by PCR), copying (such as addition by a polymerase), and non-covalent interactions.

Specific methods include including barcodes (*e.g.,* designed or random sequences) to PCR primers so that the each template will receive an individual code within the code space, yielding unique amplicons that can be discriminated from other amplicons. This concept can be applied to any method that uses polymerase amplification, such as GoldenGate assays as disclosed in U.S. Patent No. 7,582,420, No. 7,955,794, and No. 8,003,354. Code-tagged target sequences can be circularized and amplified by methods such as rolling-circle amplification to yield code-tagged amplicons. Similarly, the code can also be added to RNA

Methods for making DNA circles

The present invention also provides methods and compositions for making DNA circles by fragmenting and circularizing DNA at the same time by insertion of "hairpin-like" transposons. Previous methods of circularizing DNA are based on ligating "sticky"

IDT Exhibit 1018.0349

or blunt end ends of DNA fragments together or by the use of ligases or related enzymes to ligate 5' and 3' DNA fragments from the same template molecule.

In this embodiment, special transposons are prepared in which DNA recognition sites for the transposase are double-stranded, with the overall transposon forming a hairpin. Two transposons are loaded into a transposon-transposase complex allowing the insertion of the hairpin transposon. Since the transposon ends are linked (as in **FIG. 24**, for example), insertion will fragment the DNA and simultaneously link the top and bottom strand of the target DNA. This process can occur throughout the genome, leading to the creation of many circles of the original starting DNA. A linker that links the top and bottom strand of the transposon together and can be any standard chemical linkers or a nucleic acid linker, such as single stranded DNA or RNA. Single-stranded DNA linkers are particularly useful since any DNA sequence can be designed as a linker, allowing pre-designed DNA to be incorporated into the DNA circles. The single-stranded linker can contain primer sites for amplification purposes.

Amplification of circularized DNA can be useful for sample preparation schemes for sequencing, genotyping, genotyping of FFPE samples. The method can be used for whole genome amplification with specific primers (not randomers) of genomic DNA. For example, primers can be designed into the single-stranded linker of the transposons allowing MDA or whole genome amplification of genomic DNA. One or more primers can be designed to accomplish exponential whole genome amplification. Alternatively, the method can be used to amplify a subset of genomic DNA. The circular genomic DNA can be amplified to prepare DNA balls. Whole genome amplification can be used for sample preparation for whole genome amplification, such as for the Infinium® assay workflow (Illumina, Inc.). In a particular embodiment, the method can be used to amplify FFPE DNA for whole genome genotyping. Short FFPE (Formalin-Fixed Paraffin-Embedded) DNA does not amplify efficiently in whole genome amplification schemes (Klenow/random primers or MD-like approaches. This method can generate circular DNA and these circular templates are efficient substrates for amplification. The amplification can be combined with lesion-proof polymerases to start the amplification.

A further advantage is that pre-designed primers can be used for amplification purposes. As such, whole genome amplification can be performed with one or few primers instead of the standard randomers. This method can lead to higher uniformity of amplification of a DNA sample with less locus dropout.

IDT Exhibit 1018.0350

Targeted insertion

The method is particularly useful when tagmentation is target-specific, so that a desired subset of the genome is made into a sequencing library.

In one implementation, dsDNA is denatured into ssDNA and annealed with oligonucleotide probes (20-200 bases). These probes create sites of dsDNA that can be efficiently tagmented with the transposase-transposon complex. This approach takes advantage of the fact that transposition into dsDNA is much more efficient compared to ssDNA targets. Furthermore, the transposase complex is covalently linked to the targeting oligonucleotides so that annealing brings the transposon complex into the vicinity of the targeted region of interest. At least two targeting oligos can be employed to generate library elements (tagments) that span the region between the targeting oligos (see **FIG. 25**).

The repeat regions may quickly re-hybridize, creating tagmentation sites that are non-specific and not desirable. To prevent this from occurring, a genomic representation using non-complementary bases can be prepared. For example, the DNA circles discussed above can be prepared that are uniformly amplified with specific primers, allowing the incorporation of non-complementary or pseudo-complementary bases. The end result is a genomic DNA representation that does not interact with itself or other amplified genomic DNA molecules. Specific annealing of standard oligonucleotides to these pseudo-complementary, single-stranded, genomic DNA can create specific tagmentation sites.

In another implementation (see **FIG. 26**), dsDNA is targeted directly using oligonucleotides (or LNA) probes coupled to the transposase complex. dsDNA can be targeted using a number of approaches including D-loop formation with recA-coated oligo probes, triplex formation, and other methods described in the literature. As above, two probes should flank the region of interest to generate a final library element spanning the region between the probes. Depending on the length of the linker between the targeting oligo and the transposase complex, the complex will insert a variable distance upstream or downstream of the targeted complex.

In yet another implementation regions of interest in dsDNA can be targeted using sequence-specific DNA binding proteins such as zinc-finger complexes, and other affinity ligands to specific DNA regions. Tagmentation on either end of a targeted region

IDT Exhibit 1018.0351

creates DNA fragments that have primers on the 5' and 3' end. These primers can be used for the amplification of a specific region.

Methods for generating mate-pair libraries

The present invention also provides an improved method for preparing mate-pair libraries. Mate-pair libraries offer powerful means of mapping inserts and deletions in genomes as well as genome finishing and contig assembly. Current mate-pair library preparation methods can involve sequentially: a first fragmentation, size selection, end-repair, adapter ligation, circularization, a secondary fragmentation, junction fragment enrichment, and PCR. Though random fragmentation is desirable in general, the Poisson distribution of fragment sizes in the second fragmentation can yield numerous mate-pair arms that are too small to be of bioinformatic use. Such mate pairs must be discarded and this reduces the size of the mate-pair library.

The invention can involve tethering the circularized intermediate to the surface of a particle to use the steric hindrance of the particle to reduce access of transposomes to the DNA, reducing transposase cleavage immediately adjacent to the site of immobilization.

In the embodiment shown in **FIG. 27**, the process begins by cleaving genomic DNA with transposomes and selecting fragments of the desired size by gel purification or other fragment selection methodology. The 3'-ends of the cleavage products are then melted away and replaced with an excess quantity of a new sequence that has been amended to contain a 3'-palindrome (for subsequent circularization) and a binding residue. An exemplary binding residue is a biotin residue for enrichment on streptavidin-paramagnetic particles (SA-PMPs). In this embodiment, the newly modified fragments retain the 9-bp gaps left by the Tn5 transposase (or other enzyme that is used) and may be repaired in a gap-filling reaction driven by a non-strand displacing DNA polymerase that also lacks 3'-5' exonuclease activity, and a DNA ligase. The DNA ligase used in the gap-filling reaction can also serve to generate DNA circles. Following exonuclease-mediated removal of residual linear molecules, the circular DNAs can be captured onto the surface of a binding surface (*e.g.,* SA-PMPs) and subjected to a second tagmentation reaction. This tagmentation reaction introduces priming sites compatible with the next-generation sequencing platform of choice. In further embodiments, sequencing is performed.

IDT Exhibit 1018.0352

Since the protective effect diminishes as the distance from a tether increases, the steric hindrance reduces the number of less-desirable small mate-pair arms results in mate-pair libraries of greater complexity due to a process that precludes losses from occurring during the second cleavage reaction.

Methods of analyzing template nucleic acids

Some embodiments include methods of analyzing template nucleic acids. Sequencing information can be obtained from a template nucleic acids and a sequence representation of the target nucleic acid can be obtained from such sequencing data.

In some embodiments, a linked read strategy may be used. A linked read strategy can include identifying sequencing data that links at least two sequencing reads. For example, a first sequencing read may contain a first marker, and a second sequencing read may contain a second marker. The first and second markers can identify the sequencing data from each sequencing read to be adjacent in a sequence representation of the target nucleic acid. In some embodiments, markers can comprise a first barcode sequence and a second barcode sequence in which the first barcode sequence can be paired with the second barcode sequence. In more embodiments, markers can comprise a first host tag and a second host tag. In more embodiments, markers can comprise a first barcode sequence with a first host tag, and a second barcode sequence with a second host tag.

An exemplary embodiment of a method for sequencing a template nucleic acid can comprise the following steps. First, sequence the first barcode sequence using a primer hybridizing to the first primer site as the sequencing primer; second, sequence the second barcode sequence using a primer hybridizing to the second primer site as the sequencing primer. The result is two sequence reads that help link the read to its genomic neighbors. Given long enough reads, and short enough library fragments, these two reads can be merged informatically to make one long read that covers the entire fragment. Using the barcode sequence reads and the 9 nucleotide duplicated sequence present from the insertion, reads can now be linked to their genomic neighbors to form much longer "linked reads" *in silico*. As will be understood, a library comprising template nucleic acids can include duplicate nucleic acid fragments. Sequencing duplicate nucleic acid fragments is advantageous in methods that include creating a consensus sequence for

IDT Exhibit 1018.0353

duplicate fragments. Such methods can increase the accuracy for providing a consensus sequence for a template nucleic acid and/or library of template nucleic acids.

In some embodiments, sequence analysis is performed in real time, for example, sequence data can be obtained and simultaneously analyzed. In some embodiments, a sequencing process to obtain sequencing data can be terminated at various points, including after at least a portion of a target nucleic acid sequence data is obtained or before the entire nucleic acid read is sequenced. Exemplary methods, systems, and further embodiments are provided in International Patent Publication WO 2010/062913.

FIG. 28 depicts an exemplary embodiment of a method for assembling short sequencing reads using a linked read strategy. In such an exemplary embodiment, transposon sequences comprising barcodes are inserted into genomic DNA, a library is prepared and sequencing data obtained for the library of template nucleic acids. Blocks of templates are assembled by identifying paired barcodes and larger contigs are then assembled. **FIG. 29**, **FIG. 30** and **FIG. 31** depict exemplary embodiments of methods of assembling sequencing reads using a linked ready strategy.

Some embodiments include error detection and correction. Examples of errors can include errors in base calls during a sequencing process, and errors in assembling fragments into larger contigs. As would be understood, error detection can include detecting the presence or likelihood of errors in a data set, and as such detecting the location of an error or number of errors may not be required. For error correction, information regarding the location of an error and/or the number of errors in a data set is useful. Methods for error correction are well known in the art. Examples include the use of hamming distances, and the use of a checksum algorithm (*See, e.g.,*U.S. Publication 2010/0323348; U.S. Patent No. 7,574,305; U.S. Patent No. 6,654,696).

— nested libraries

An alternative method involves the junction tagging methods above and preparation of nested sequencing libraries, as illustrated in **FIG. 32**. The nested sub-libraries are created from code-tagged DNA fragments. This can allow less frequent transposon tagging across the genome. It can also create a larger diversity of (nested) sequencing reads. These factors can lead to improved coverage and accuracy.

Sub-sampling and whole genome amplification can create many copies of a certain population of starting molecules. DNA fragments are then generated by

IDT Exhibit 1018.0354

transposon-specific fragmentation, where each fragment receives a code that allows one to link the fragment back to the original neighbor having a matching code (whether identical, complementary or otherwise informatically linked). The tagged fragments are fragmented at least a second time by methods random or sequence-specific methods such as enzymatic digestion, random shearing, transposon-based shearing or other methods, thereby creating sub-libraries of the code-tagged DNA fragments. In a useful variation, code-tagged fragments can be preferentially isolated by using transposons that contain a biotin or other affinity functionality for downstream enrichment purposes. Subsequently, library preparation converts the nested DNA fragments into sequencing templates. Paired-end sequencing results in sequencing the of the code-tag of the DNA fragments and of the target DNA. Since nested libraries for the same code-tag are created, long DNA fragments can be sequenced with short reads.

— two-step sub-assembly

In another embodiment, a two-step tagging method is provided using one or more transposition reactions with different transposomes. The transposomes can be of different types (e.g. Mu versus Tn5) or be two sets of Tn5 transposomes with different sequences, for example.

As exemplified in **FIG. 33**, a first transposome can use a Mu-type transposase, such as HyperMu™ (Epicentre). The DNA section of this complex is barcoded at the individual molecule level and presence of a molecular barcode at both ends of each fragment ("end-tags"). Each DNA fragment receives a unique tag (code-tagging) for the purpose of making each DNA fragment unique. The library is then amplified via PCR, so that every fragment is represented multiple times in the amplified reaction. The amplified library is next subjected to a second round of tagmentation with a different transposome, such as Tn5. If desired, the concentration of enzyme can be limited so every molecule is minimally targeted with the second transposome to preserve maximum continuity information. In addition, complexity reduction schemes at various stages throughout the assay can be utilized to ensure that the majority of libraries sequenced cover a significant fraction of the target DNA.

IDT Exhibit 1018.0355

— compartmentalization of single molecules

Some embodiments include the use of the compositions and methods provided herein in single cell applications. In some embodiments, a transposon sequence can be associated with a bead via a cleavable linker. A bead comprising the transposon sequence and transposase is provided in a droplet. The cell is lysed and the transposon sequence cleaved from the bead. The transposition reaction is initiated. Such methods can be used to provide a library of template nucleic acids for each cell in population. Methods particularly useful for single-cell applications include clonal transposon beads, transposase mix, and single cell/single DNA molecule suspensions.

In an embodiment illustrated in **FIG. 34**, droplets or microwells are used to compartmentalize a single transposon bead with a single cell/DNA molecule in the presence of tagmentation reaction mix. The transposon is immobilized to the bead through a cleavable linker, such as a photocleavable linker. Each bead has a unique transposon index sequence for the appropriate barcoding of a particular cell. Additionally, photo-caged Mg^{++} can be used to activate the tagmentation reaction at the same time the oligos are cleaved from the bead. Upon exposure to an appropriate wavelength of light (*e.g.*, 365 nm for many photoreactive compounds), the transposon is cleaved from the bead and assembles with the transposase to create an active complex. In the presence of Mg^{++}, this complex will insert into the available target DNA and fragment the sample. The DNA sample now contains a unique index along with adapter sequences from the transposon. All the compartmentalized reactions are simultaneously lysed/extracted into a single eluant. This eluant can be amplified with adapter-primers to create final sequencing library. For methylation analysis, the eluant can be treated with bisulfite first, and then amplified to enable bisulfite sequence analysis.

In an embodiment exemplified in **FIG. 35,** beads with uniquely indexed transposons (attached or hybridized via identifier sequences) are loaded into a microwell to create a transposon bead array. Accordingly, the present invention provides a transposon bead array. The transposon beads can be created in a number of ways. One way is clonal amplification of single-transposon seed molecules using bridge amplification of emulsion PCR. Another way is split-and-pool oligonucleotide synthesis in which the index portion of the transposon on the bead is generated during a split-and-pool oligonucleotide bead synthesis. A further way is to hybridize indexed transposons with an address sequence to beads containing the cognate address sequence.

IDT Exhibit 1018.0356

Alternatively, the transposon bead array can be preloaded with beads having index sequences and then hybridizing indexed transposons. The transposon bead array is loaded with tagmentation mix and cells/DNA so that each well has less than one cell on average. The transposon bead array can be overlaid with oil and heated to lyse the cells. The transposons are released from the bead via photocleavage. Alternatively, photo-caged Mg^{++} can be released to activate the tagmentation reaction. After tagmentation, the wells are extracted and the eluate is pooled. The eluate can be amplified (such as by PCR) to yield a cell-indexed sequencing library.

Sequencing methods

The methods described herein can be used in conjunction with a variety of sequencing techniques. In some embodiments, the process to determine the nucleotide sequence of a target nucleic acid can be an automated process.

Some embodiments include pyrosequencing techniques. Pyrosequencing detects the release of inorganic pyrophosphate (PP_i) as particular nucleotides are incorporated into the nascent strand (Ronaghi *et al.*, "Real-time DNA sequencing using detection of pyrophosphate release." *Analytical Biochemistry* 242(1): 84-9 (1996); Ronaghi, M. "Pyrosequencing sheds light on DNA sequencing" *Genome Res.* 11(1):3-11 (2001); Ronaghi *et al.*, "A sequencing method based on real-time pyrophosphate." *Science* 281(5375):363 (1998); U.S. Patent No. 6,210,891; U.S. Patent No. 6,258,568 and U.S. Patent No. 6,274,320). In pyrosequencing, released PP_i can be detected by being immediately converted to adenosine triphosphate (ATP) by ATP sulfurylase, and the level of ATP generated is detected via luciferase-produced photons.

In another example type of SBS, cycle sequencing is accomplished by stepwise addition of reversible terminator nucleotides containing, for example, a cleavable or photobleachable dye label as described, for example, in U.S. Patent No. 7,427,67, U.S. Patent No. 7,414,1163 and U.S. Patent No. 7,057,026. This approach is being commercialized by Solexa (now Illumina Inc.), and is also described in WO 91/06678 and WO 07/123744 (filed in the United States patent and trademark Office as US 12/295,337). The availability of fluorescently-labeled terminators in which both the termination can be reversed and the fluorescent label cleaved facilitates efficient cyclic reversible termination (CRT) sequencing. Polymerases can also be co-engineered to efficiently incorporate and extend from these modified nucleotides.

IDT Exhibit 1018.0357

Additional example SBS systems and methods which can be utilized with the methods and systems described herein are described in U.S. Patent Application Publication No. 2007/0166705, U.S. Patent Application Publication No. 2006/0188901, U.S. Patent No. 7057026, U.S. Patent Application Publication No. 2006/0240439, U.S. Patent Application Publication No. 2006/0281109, PCT Publication No. WO 05/065814, U.S. Patent Application Publication No. 2005/0100900, PCT Publication No. WO 06/064199 and PCT Publication No. WO 07/010251.

Some embodiments can utilize sequencing by ligation techniques. Such techniques utilize DNA ligase to incorporate nucleotides and identify the incorporation of such nucleotides. Example SBS systems and methods which can be utilized with the methods and systems described herein are described in U.S. Patent No 6,969,488, U.S. Patent No. 6,172,218, and U.S. Patent No. 6,306,597.

Some embodiments can include techniques such as next-next technologies. One example can include nanopore sequencing techniques (Deamer, D.W. & Akeson, M. "Nanopores and nucleic acids: prospects for ultrarapid sequencing." *Trends Biotechnol.* 18, 147–151 (2000); Deamer, D. and D. Branton, "Characterization of nucleic acids by nanopore analysis". *Acc. Chem. Res.* 35:817-825 (2002); Li *et al.,* "DNA molecules and configurations in a solid-state nanopore microscope" *Nat. Mater.* 2:611–615 (2003)). In such embodiments, the target nucleic acid passes through a nanopore. The nanopore can be a synthetic pore or biological membrane protein, such as α-hemolysin. As the target nucleic acid passes through the nanopore, each base-pair can be identified by measuring fluctuations in the electrical conductance of the pore. (U.S. Patent No. 7,001,792; Soni & Meller, "A. Progress toward ultrafast DNA sequencing using solid-state nanopores." *Clin. Chem.* 53, 1996–2001 (2007); Healy, K. "Nanopore-based single-molecule DNA analysis." *Nanomed.* 2:459–481 (2007); Cockroft *et al.,* "A single-molecule nanopore device detects DNA polymerase activity with single-nucleotide resolution." *J. Am. Chem. Soc.* 130:818–820 (2008)). In some such embodiments, nanopore sequencing techniques can be useful to confirm sequence information generated by the methods described herein.

Some embodiments can utilize methods involving the real-time monitoring of DNA polymerase activity. Nucleotide incorporations can be detected through fluorescence resonance energy transfer (FRET) interactions between a fluorophore-bearing polymerase and γ-phosphate-labeled nucleotides as described, for example, in U.S. Patent No. 7,329,492 and U.S. Patent No. 7,211,414 or nucleotide incorporations can

IDT Exhibit 1018.0358

be detected with zero-mode waveguides as described, for example, in U.S. Patent No. 7,315,019 and using fluorescent nucleotide analogs and engineered polymerases as described, for example, in U.S. Patent No. 7,405,281 and U.S. Patent Application Publication No. 2008/0108082 The illumination can be restricted to a zeptoliter-scale volume around a surface-tethered polymerase such that incorporation of fluorescently labeled nucleotides can be observed with low background (Levene, M.J. *et al.* "Zero-mode waveguides for single-molecule analysis at high concentrations." *Science* 299, 682–686 (2003); Lundquist, P.M. *et al.* "Parallel confocal detection of single molecules in real time." *Opt. Lett.* 33, 1026–1028 (2008); Korlach, J. *et al.* "Selective aluminum passivation for targeted immobilization of single DNA polymerase molecules in zero-mode waveguide nanostructures." *Proc. Natl. Acad. Sci. USA* 105, 1176–1181 (2008)). In one example single molecule, real-time (SMRT) DNA sequencing technology provided by Pacific Biosciences Inc can be utilized with the methods described herein. In some embodiments, a SMRT chip or the like may be utilized (U.S. Patent Nos. 7,181,122, 7,302,146, 7,313,308). A SMRT chip comprises a plurality of zero-mode waveguides (ZMW). Each ZMW comprises a cylindrical hole tens of nanometers in diameter perforating a thin metal film supported by a transparent substrate. When the ZMW is illuminated through the transparent substrate, attenuated light may penetrate the lower 20-30 nm of each ZMW creating a detection volume of about 1×10^{-21} L. Smaller detection volumes increase the sensitivity of detecting fluorescent signals by reducing the amount of background that can be observed.

SMRT chips and similar technology can be used in association with nucleotide monomers fluorescently labeled on the terminal phosphate of the nucleotide (Korlach J. *et al.,* "Long, processive enzymatic DNA synthesis using 100% dye-labeled terminal phosphate-linked nucleotides." Nucleosides, Nucleotides and Nucleic Acids, 27:1072–1083, 2008). The label is cleaved from the nucleotide monomer on incorporation of the nucleotide into the polynucleotide. Accordingly, the label is not incorporated into the polynucleotide, increasing the signal: background ratio. Moreover, the need for conditions to cleave a label from labeled nucleotide monomers is reduced.

An additional example of a sequencing platform that may be used in association with some of the embodiments described herein is provided by Helicos Biosciences Corp. In some embodiments, TRUE SINGLE MOLECULE SEQUENCING can be utilized (Harris T.D. *et al.,* "Single Molecule DNA Sequencing of a viral Genome" *Science*

IDT Exhibit 1018.0359

320:106-109 (2008)). In one embodiment, a library of target nucleic acids can be prepared by the addition of a 3′ poly(A) tail to each target nucleic acid. The poly(A) tail hybridizes to poly(T) oligonucleotides anchored on a glass cover slip. The poly(T) oligonucleotide can be used as a primer for the extension of a polynucleotide complementary to the target nucleic acid. In one embodiment, fluorescently-labeled nucleotide monomer, namely, A, C, G, or T, are delivered one at a time to the target nucleic acid in the presence DNA polymerase. Incorporation of a labeled nucleotide into the polynucleotide complementary to the target nucleic acid is detected, and the position of the fluorescent signal on the glass cover slip indicates the molecule that has been extended. The fluorescent label is removed before the next nucleotide is added to continue the sequencing cycle. Tracking nucleotide incorporation in each polynucleotide strand can provide sequence information for each individual target nucleic acid.

An additional example of a sequencing platform that can be used in association with the methods described herein is provided by Complete Genomics Inc. Libraries of target nucleic acids can be prepared where target nucleic acid sequences are interspersed approximately every 20 bp with adaptor sequences. The target nucleic acids can be amplified using rolling circle replication, and the amplified target nucleic acids can be used to prepare an array of target nucleic acids. Methods of sequencing such arrays include sequencing by ligation, in particular, sequencing by combinatorial probe-anchor ligation (cPAL).

In some embodiments using cPAL, about 10 contiguous bases adjacent to an adaptor may be determined. A pool of probes that includes four distinct labels for each base (A, C, T, G) is used to read the positions adjacent to each adaptor. A separate pool is used to read each position. A pool of probes and an anchor specific to a particular adaptor is delivered to the target nucleic acid in the presence of ligase. The anchor hybridizes to the adaptor, and a probe hybridizes to the target nucleic acid adjacent to the adaptor. The anchor and probe are ligated to one another. The hybridization is detected and the anchor-probe complex is removed. A different anchor and pool of probes is delivered to the target nucleic acid in the presence of ligase.

The sequencing methods described herein can be advantageously carried out in multiplex formats such that multiple different target nucleic acids are manipulated simultaneously. In particular embodiments, different target nucleic acids can be treated in a common reaction vessel or on a surface of a particular substrate. This allows convenient

IDT Exhibit 1018.0360

delivery of sequencing reagents, removal of unreacted reagents and detection of incorporation events in a multiplex manner. In embodiments using surface-bound target nucleic acids, the target nucleic acids can be in an array format. In an array format, the target nucleic acids can be typically coupled to a surface in a spatially distinguishable manner. For example, the target nucleic acids can be bound by direct covalent attachment, attachment to a bead or other particle or associated with a polymerase or other molecule that is attached to the surface. The array can include a single copy of a target nucleic acid at each site (also referred to as a feature) or multiple copies having the same sequence can be present at each site or feature. Multiple copies can be produced by amplification methods such as, bridge amplification or emulsion PCR as described in further detail herein.

The methods set forth herein can use arrays having features at any of a variety of densities including, for example, at least about 10 features/cm^2, 100 features/cm^2, 500 features/cm^2, 1,000 features/cm^2, 5,000 features/cm^2, 10,000 features/cm^2, 50,000 features/cm^2, 100,000 features/cm^2, 1,000,000 features/cm^2, 5,000,000 features/cm^2, 10^7 features/cm^2, 5x 10^7 features/cm^2, 10^8 features/cm^2, 5x 10^8 features/cm^2, 10^9 features/cm^2, 5x 10^9 features/cm^2, or higher.

Surfaces

In some embodiments, the nucleic acid template provided herein can be attached to a solid support ("substrate"). Substrates can be two-or three-dimensional and can comprise a planar surface (*e.g.,* a glass slide) or can be shaped. A substrate can include glass (*e.g.,* controlled pore glass (CPG)), quartz, plastic (such as polystyrene (low cross-linked and high cross-linked polystyrene), polycarbonate, polypropylene and poly(methylmethacrylate)), acrylic copolymer, polyamide, silicon, metal (*e.g.,* alkanethiolate-derivatized gold), cellulose, nylon, latex, dextran, gel matrix (*e.g.,* silica gel), polyacrolein, or composites.

Suitable three-dimensional substrates include, for example, spheres, microparticles, beads, membranes, slides, plates, micromachined chips, tubes (*e.g.,* capillary tubes), microwells, microfluidic devices, channels, filters, or any other structure suitable for anchoring a nucleic acid. Substrates can include planar arrays or matrices capable of having regions that include populations of template nucleic acids or primers.

IDT Exhibit 1018.0361

Examples include nucleoside-derivatized CPG and polystyrene slides; derivatized magnetic slides; polystyrene grafted with polyethylene glycol, and the like.

Various methods can be used to attach, anchor or immobilize nucleic acids to the surface of the substrate. The immobilization can be achieved through direct or indirect bonding to the surface. The bonding can be by covalent linkage. *See* Joos *et al.* (1997) Analytical Biochemistry, 247:96-101; Oroskar *et al.* (1996) Clin. Chem., 42:1547-1555; and Khandjian (1986) Mol. Bio. Rep., 11:107-11. A preferred attachment is direct amine bonding of a terminal nucleotide of the template or the primer to an epoxide integrated on the surface. The bonding also can be through non-covalent linkage. For example, biotin-streptavidin (Taylor *et al.* (1991) J. Phys. D: Appl. Phys., 24:1443,) and digoxigenin with anti-digoxigenin (Smith *et al., Science*, 253:1122 (1992)), are common tools for anchoring nucleic acids to surfaces and parallels. Alternatively, the attachment can be achieved by anchoring a hydrophobic chain into a lipid monolayer or bilayer. Other methods known in the art for attaching nucleic acid molecules to substrates can also be used.

The following Examples provide illustrative embodiments and do not in any way limit the inventions provided herein.

EXAMPLES

Example 1—Whole genome amplification using transposon sequences

This example illustrates a method for uniform amplification of genomic DNA with random insertion therein of specific primer sites. Transposon sequences are prepared, each comprising a first transposase recognition site, a second transposase recognition site having a sequencing adaptor disposed therebetween, in which the sequencing adaptor comprises a first primer site and second primer site. The transposon sequences are contacted with genomic DNA in the presence of MuA transposase under conditions sufficient for the transposon sequences to integrate into the genomic DNA. The genomic DNA is amplified using primers that hybridize to the first primer site or second primer site.

IDT Exhibit 1018.0362

Example 2—Landmark sequencing methods using genomes with increased complexity

This example illustrates an embodiment for providing additional markers in a genome. Additional markers can be useful in genomes that include repetitive sequences during subsequent assembly steps to generate a sequence representation of the genome. Transposon sequences are prepared, each comprising a different barcode. The transposon sequences are integrated into genomic DNA in a transposition reaction. The genomic DNA comprising the integrated transposon is amplified by whole genome amplification. A sequencing library is prepared from the amplified template nucleic acids. Sequencing data is obtained from the sequencing library. Sequencing reads can include representations of one or more nucleic acids with the same barcode on each nucleic acid. Such nucleic acids are identified as containing sequences that overlap in a sequence representation of the genomic DNA. The sequencing reads can be assembled by identifying barcodes on overlapping sequences.

Example 3—Predicted average coverage using linked read sequencing strategy

Useable fragment lengths are modeled as a truncated exponential distribution so that the mean useable fragment length can be obtained by setting k = b/d, where d is the mean of the non-truncated exponential (the total fragment distribution) and b is the value for truncation (either 180 or 280 for 100 nucleotides and 150 nucleotide paired-end reads, respectively) and then calculating the mean of the truncated exponential as

$$E(f) = d\,(1-(k+1)e^{-k}) / (1 - e^{-k})$$

The proportion of useable reads is p = C(b) x (1 - D(0,T)) where C is the exponential cumulative distribution function, T is the average repetitions of observing each fragment (num clusters) / complexity, complexity is the genome size times the number of genome copies diluted to divided by d, and D is the Poisson cumulative distribution function

Expected length of linked read is then (E(f) – 9) x 1 / (1 – p) + 9 where p is proportion of useable reads: 9 is subtracted from each read because of the reused 9 nucleotides sequence in neighboring fragments; and 9 nucleotides is added back to the linked read to account for the end read –one of its 9 nucleotides segments is not shared with another read within the linked read.

IDT Exhibit 1018.0363

The distribution of linked read lengths is also exponential with the above expected value. Very long linked reads can be observed, and shown in **FIG. 36**.

Table 1 sets out predicted average coverage for a bacterial genome, human PCR product, and a single copy of a human diploid genome using a linked read strategy.

TABLE 1

Target nucleic acid	Genome size	Read length	Diluted copies	Number of clusters	Average insert distance	Average coverage	Mean linked read length
Bacterial genome	5.0e6	150	10	35,000,000	50 nt	9.96 X	10,815 nt
Human PCR product	20 kb	150	10,000	40,000,000	50 nt	9.96 X	10,811 nt
Single cell, human diploid genome	3e9	150	2	3000,000,000	50 nt	9.66 X	1,191 nt

Example 4—*De novo* sequencing target DNA

This example illustrates an embodiment of assembling sequencing data obtained from a library of template nucleic acids prepared from a target DNA.

A plurality of transposon sequences are integrated into the target DNA in a transposition reaction. Each transposon sequence includes a barcode which comprises a first barcode sequence and a second barcode sequence. The first barcode sequence is the reverse complement of the second barcode sequence. There are more than 10^{18} different barcodes in the plurality of transposon sequences, such that the code space is large and out-competes the complexity of the target DNA. Accordingly, each integrated barcode is likely to be unique. Optionally, the target DNA comprising the integrated transposon sequences is sub-sampled. The sub-sampled target DNA comprising the integrated transposon sequences is amplified by methods of whole genome amplification. A sequencing library is prepared from the amplified nucleic acids. Sequencing data is obtained from the sequencing library. The sequencing data comprises sequencing reads for each amplified nucleic acid.

IDT Exhibit 1018.0364

Sequencing reads that include a barcode sequence that can be paired with a barcode sequence of another sequencing read are aligned. Shorter alignments of sequencing reads are aligned with other short alignments by identifying paired barcode sequences to produce longer alignments. A sequence representation of the target DNA is generated.

Example 5—Preparing a linked library with blunt-end barcodes

This example illustrates an embodiment of preparing a linked library that includes identical barcodes on template nucleic acids that include adjacent sequences in a sequence representation of the target nucleic acid.

Transposon sequences are prepared comprising a first fragmentation site and a second fragmentation site, having a barcode disposed therebetween. Each fragmentation site comprises a site that can be nicked to produce a single-stranded sticky end, *e.g.*, a restriction endonuclease site that produces a single-strand sticky end. The transposon sequences are integrated into the target DNA by a transposition reaction. Optionally, the target DNA comprising the integrated transposon sequences is sub-sampled. The sub-sampled target DNA comprising the integrated transposon sequences is amplified by methods of whole genome amplification. The amplified nucleic acids are fragmented at the first and second fragmentation sites at each integrated transposon sequence to generate nucleic acids comprising sticky ends. The sticky ends are filled-in so that each end comprises a barcode. Adaptors are ligated to the blunt ends of each nucleic acid. The nucleic acids are amplified using primer sites of the adaptors.

A sequencing library is prepared from the amplified nucleic acids. Sequencing data is obtained from the sequencing library. The sequencing data comprises sequencing reads for each amplified nucleic acid. Sequencing reads that include a barcode sequence that can be paired with a barcode sequence of another sequencing read are aligned. Shorter alignments of sequencing reads are aligned with other short alignments by identifying paired barcode sequences to produce longer alignments. A sequence representation of the target DNA is generated.

FIG. 37 depicts the insertion of the transposon sequence into the target nucleic acid (host material), dilution of the template nucleic acids, and subsequent whole genome amplification of the template nucleic acids. The amplified template nucleic acids are fragmented in the presence of a restriction endonuclease that nicks the transposon

IDT Exhibit 1018.0365

sequences at the first restriction endonuclease site and the second restriction endonuclease site to yield two sticky ends.

FIG. 15 depicts a fill-in reaction of the sticky ends, followed by A-tailing the ends and adding adaptors to the tailed ends. The adaptors can be used in subsequent amplification, library preparation, and methods to obtain sequence data. In other embodiments, the fill-in reaction itself can be used as part of a detection assay, such as by incorporating detectably labeled nucleotides in a variety of assay formats. Advantageously, methods that utilize a nicking endonuclease may be carried out with a relatively small number of target nucleic acids, or a relatively dilute concentration of the target nucleic acid.

The above description discloses several methods and systems of the present invention. This invention is susceptible to modifications in the methods and materials, as well as alterations in the fabrication methods and equipment. Such modifications will become apparent to those skilled in the art from a consideration of this disclosure or practice of the invention disclosed herein. For example, the invention has been exemplified using nucleic acids but can be applied to other polymers as well. Consequently, it is not intended that this invention be limited to the specific embodiments disclosed herein, but that it cover all modifications and alternatives coming within the true scope and spirit of the invention.

All references cited herein including, but not limited to, published and unpublished applications, patents, and literature references, are incorporated herein by reference in their entirety and are hereby made a part of this specification. To the extent publications and patents or patent applications incorporated by reference contradict the disclosure contained in the specification, the specification is intended to supersede and/or take precedence over any such contradictory material.

The term "comprising" as used herein is synonymous with "including," "containing," or "characterized by," and is inclusive or open-ended and does not exclude additional, unrecited elements or method steps.

IDT Exhibit 1018.0366

WHAT IS CLAIMED IS:

1. An artificial transposon, comprising:

a first transposase recognition site,

a second transposase recognition site, and

a barcode disposed therebetween, wherein the barcode comprises a double-stranded nucleic acid sequence comprising

a first strand barcode and

a second strand barcode.

2. The transposon of claim 1, wherein the transposase is a hyperactive Tn5 transposase.

3. The transposon of claim 1, wherein the transposase is a Mu transposase.

4. The transposon of claim 1, wherein the transposase is an IS5 or an IS911 transposase.

5. The transposon of claim 1, wherein the transposon is associated with an affinity tag.

6. The transposon of claim 5, wherein the affinity tag is associated with a protein that targets specific nucleic acid sequences.

7. The transposon of claim 1, wherein the first strand barcode and the second strand barcode comprise complementary sequences.

8. The transposon of claim 1, wherein the first strand barcode and the second strand barcode comprise non-complementary sequences.

9. The transposon of claim 1, further comprising at least one universal primer site.

IDT Exhibit 1018.0367

10. An isolated template nucleic acid, comprising at least a portion of a target nucleic acid or copy thereof and at least first and second transposons of claim 1, wherein the barcode of the first transposon is different from the barcode of the second transposon.

11. A method for preparing a library of template nucleic acids comprising:
(a) contacting a target nucleic acid with a plurality of transposon of claim 1 under conditions such that a portion of said of transposons are inserted into the target nucleic acid; and
(b) fragmenting said target nucleic acid.

12. The method of claim 11, wherein the target nucleic acid comprises cDNAs from a single cell.

13. The method of claim 11, wherein the target nucleic acid comprises nucleic acids from multiple species.

14. The method of claim 11, wherein the target nucleic acid comprises nucleic acids from multiple haplotypes.

15. The method of claim 11, wherein the fragmentation comprises contacting the target nucleic acid with a polymerase.

16. The method of claim 15, wherein the polymerase has 3' to 5' exonuclease activity.

17. A method for preparing a library of template nucleic acids, comprising:
(a) contacting a target nucleic acid with a plurality of transposons of claim 1, further comprising a first primer site, under conditions such that a portion of said transposons are inserted into the target nucleic acid; and
(b) hybridizing a primer to the first primer site; and
(c) amplifying a portion of the transposon sequence.

18. A method of analyzing a target nucleic acid comprising:

IDT Exhibit 1018.0368

(a) performing the method of claim 17;

(b) obtaining sequence data from said templates; and

(c) assembling a representation of a portion of said target nucleic acid from said sequence data.

19. The method of claim 18, wherein assembling step (c) comprises identifying the presence of a first barcode in one sequencing read and the presence of a corresponding barcode in another sequencing read, thereby indicating proximity between the two sequencing reads in the target nucleic acid.

20. The method of claim 13, wherein step (b) is terminated after the barcode is identified.

IDT Exhibit 1018.0369

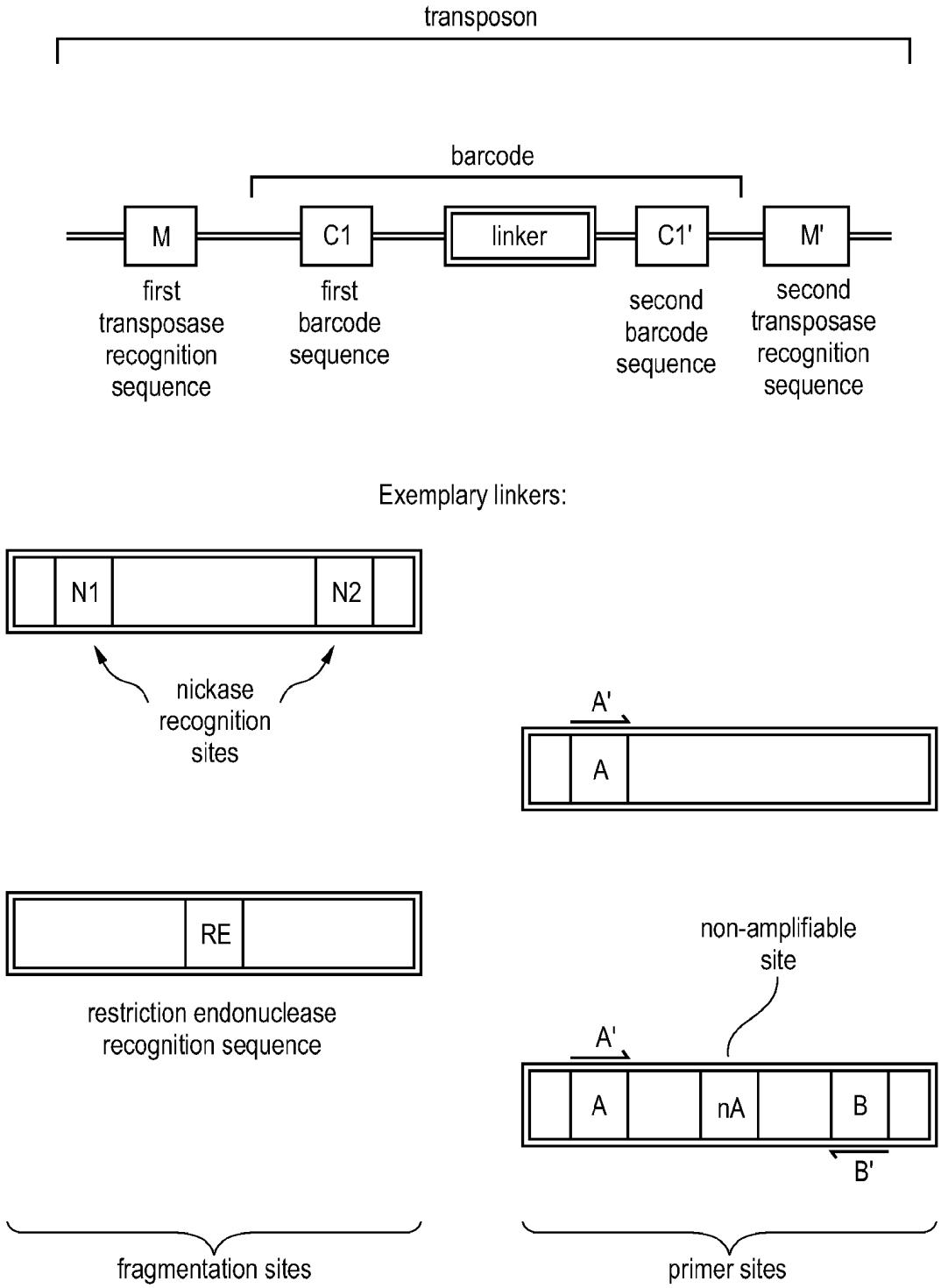


Fig. 1

IDT Exhibit 1018.0370

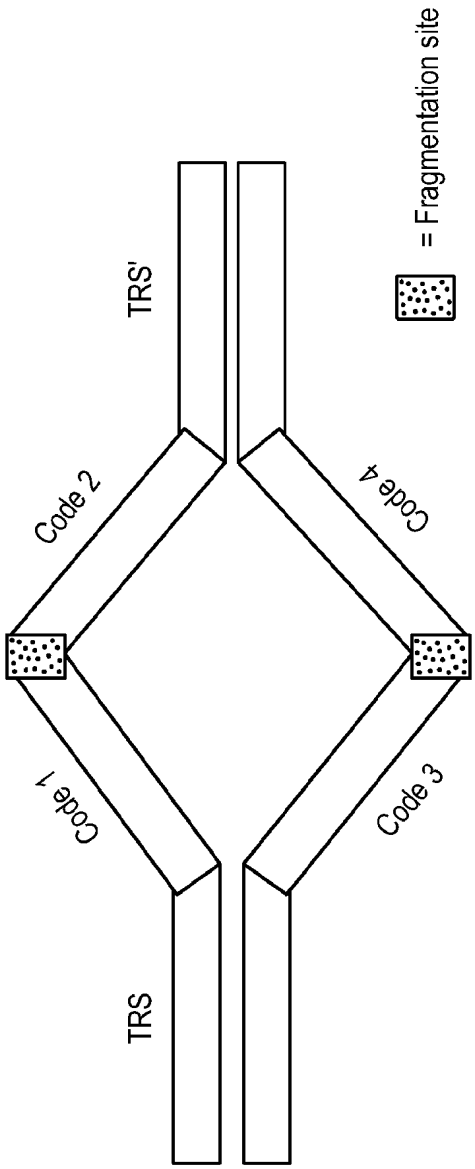


Fig. 2

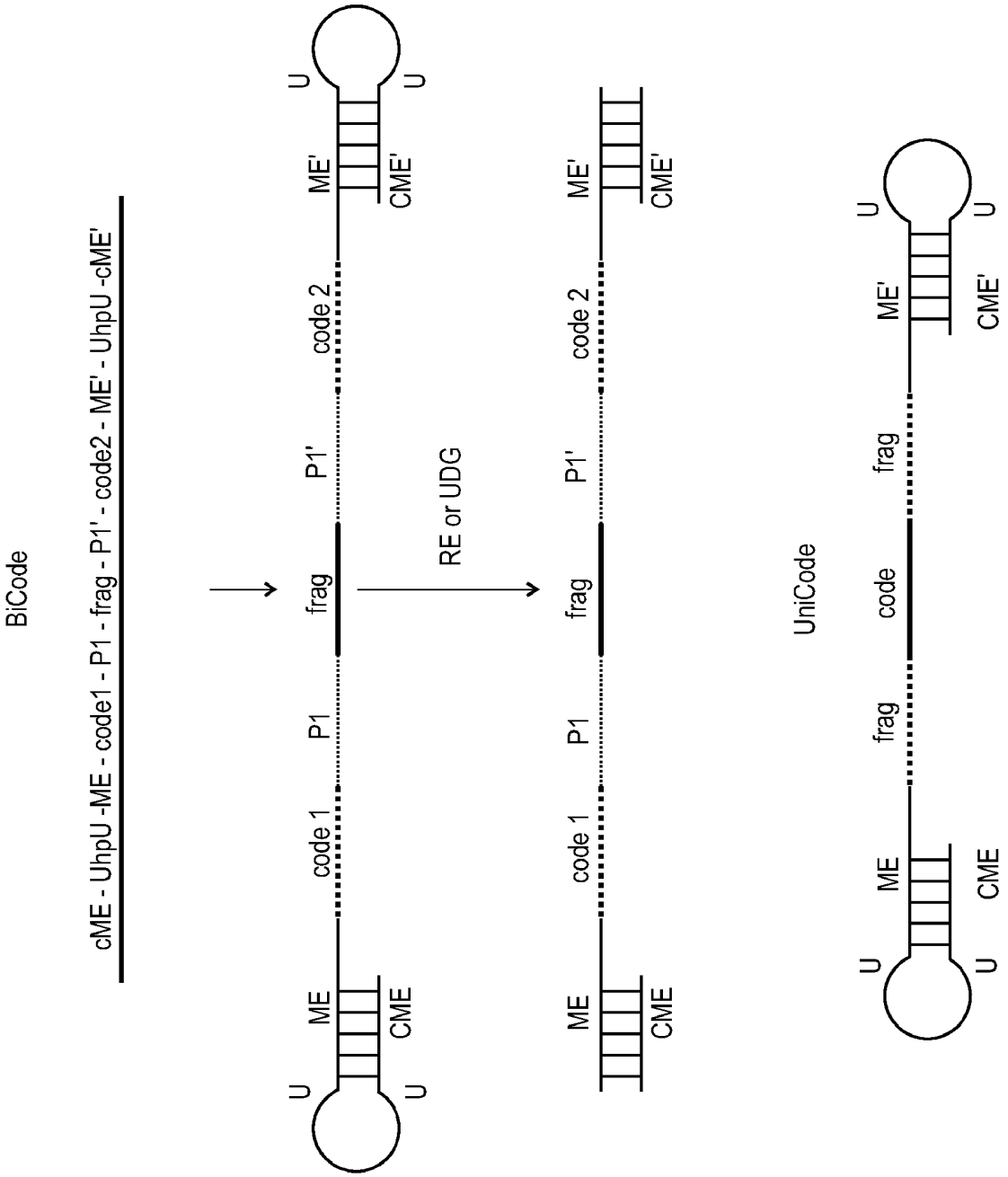

BiCode
cME - UhpU -ME - code1 - P1 - frag - P1' - code2 - ME' - UhpU -cME'
U
ME
CME
U
code 1
P1
frag
P1'
code 2
ME'
CME'
U
RE or UDG
UniCode
U
ME
CME
U
frag
code
frag
U
ME'
CME'
U

Fig. 3

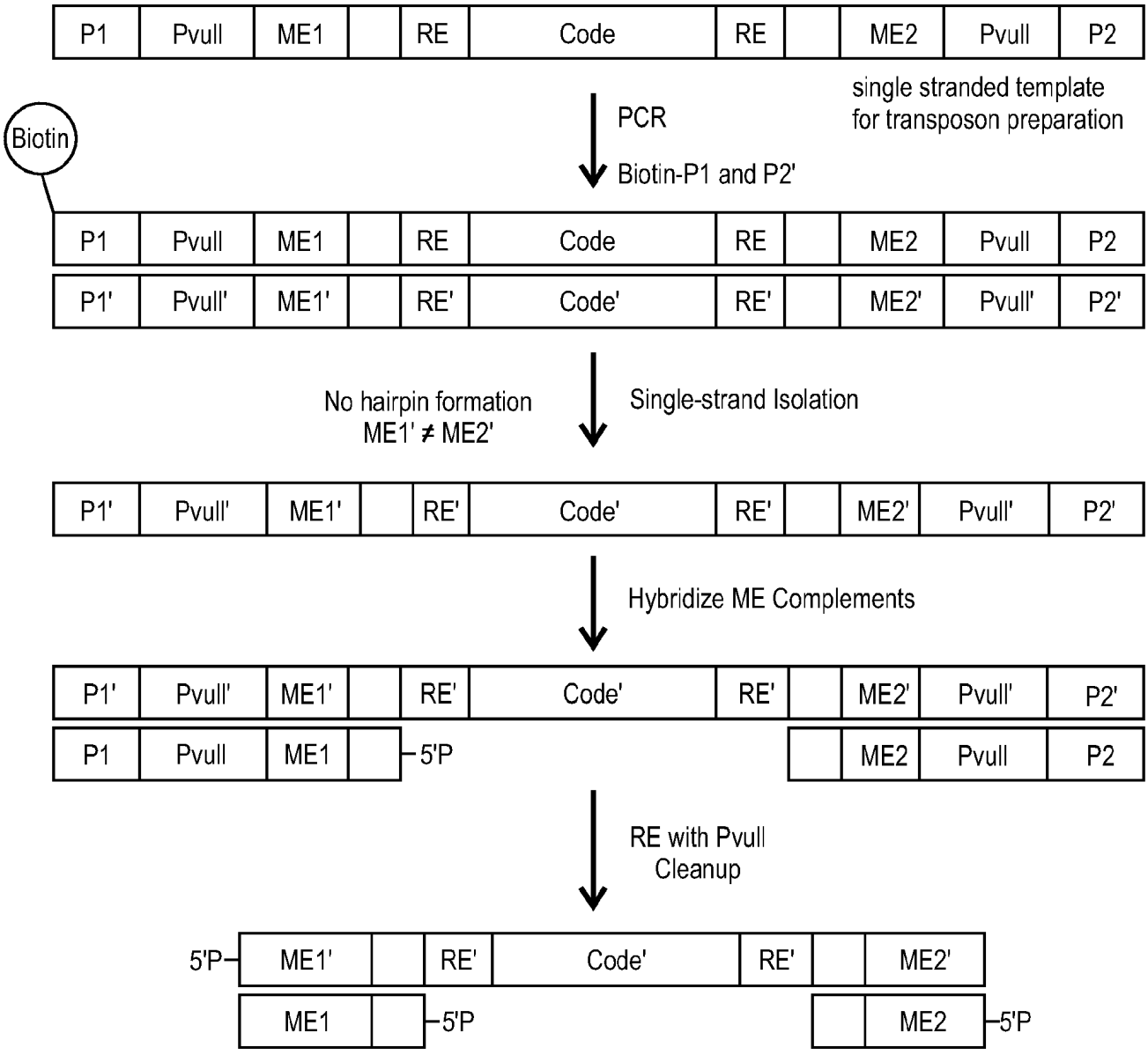


Fig. 4

IDT Exhibit 1018.0373

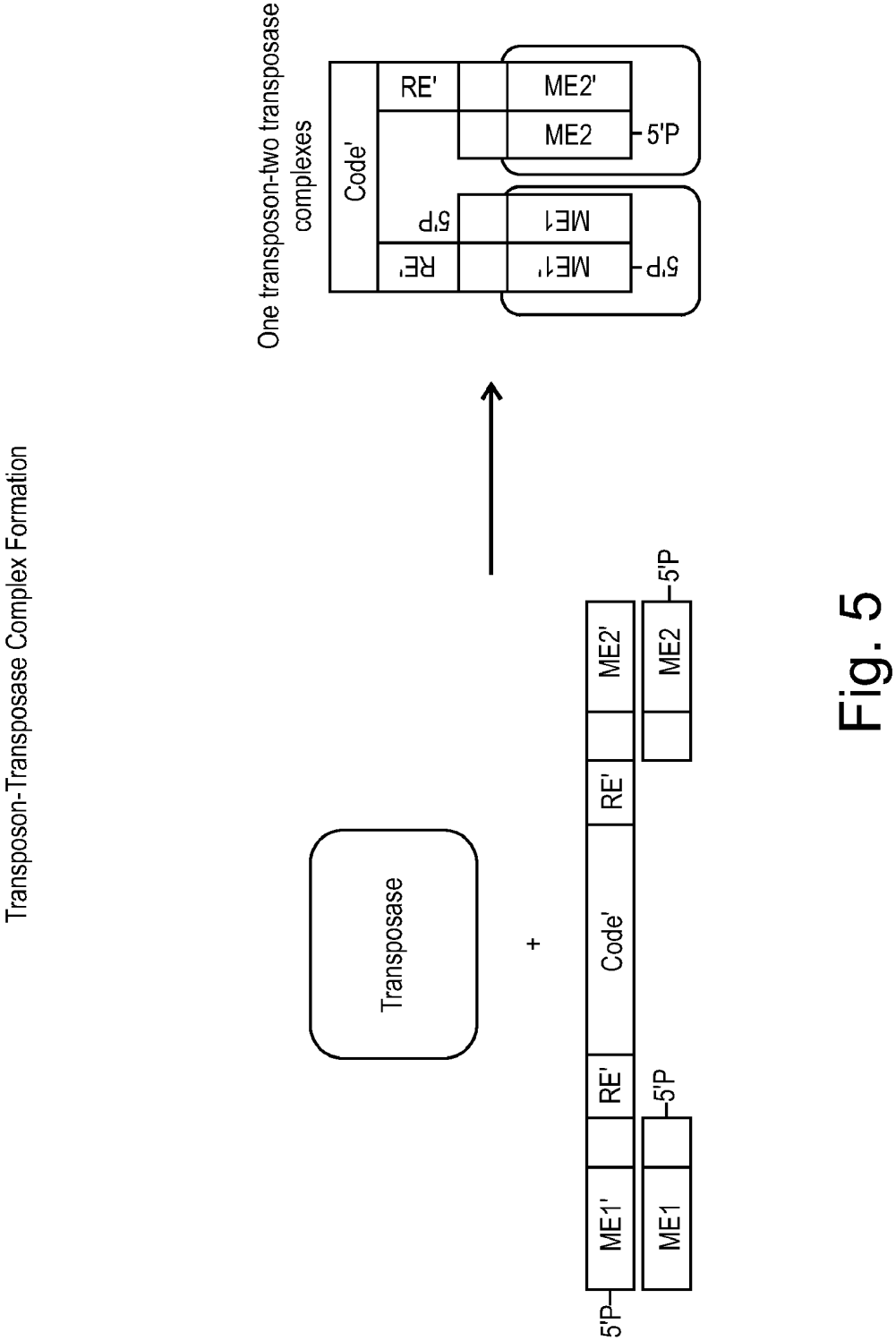

Transposon-Transposase Complex Formation
Transposase
+
5'P
ME1'
RE'
Code'
RE'
ME2'
5'P
ME1
ME2
5'P
One transposon-two transposase complexes
Code'
RE'
ME2'
ME2
5'P
5'P
RE'
ME1'
5'P
ME1

Fig. 5

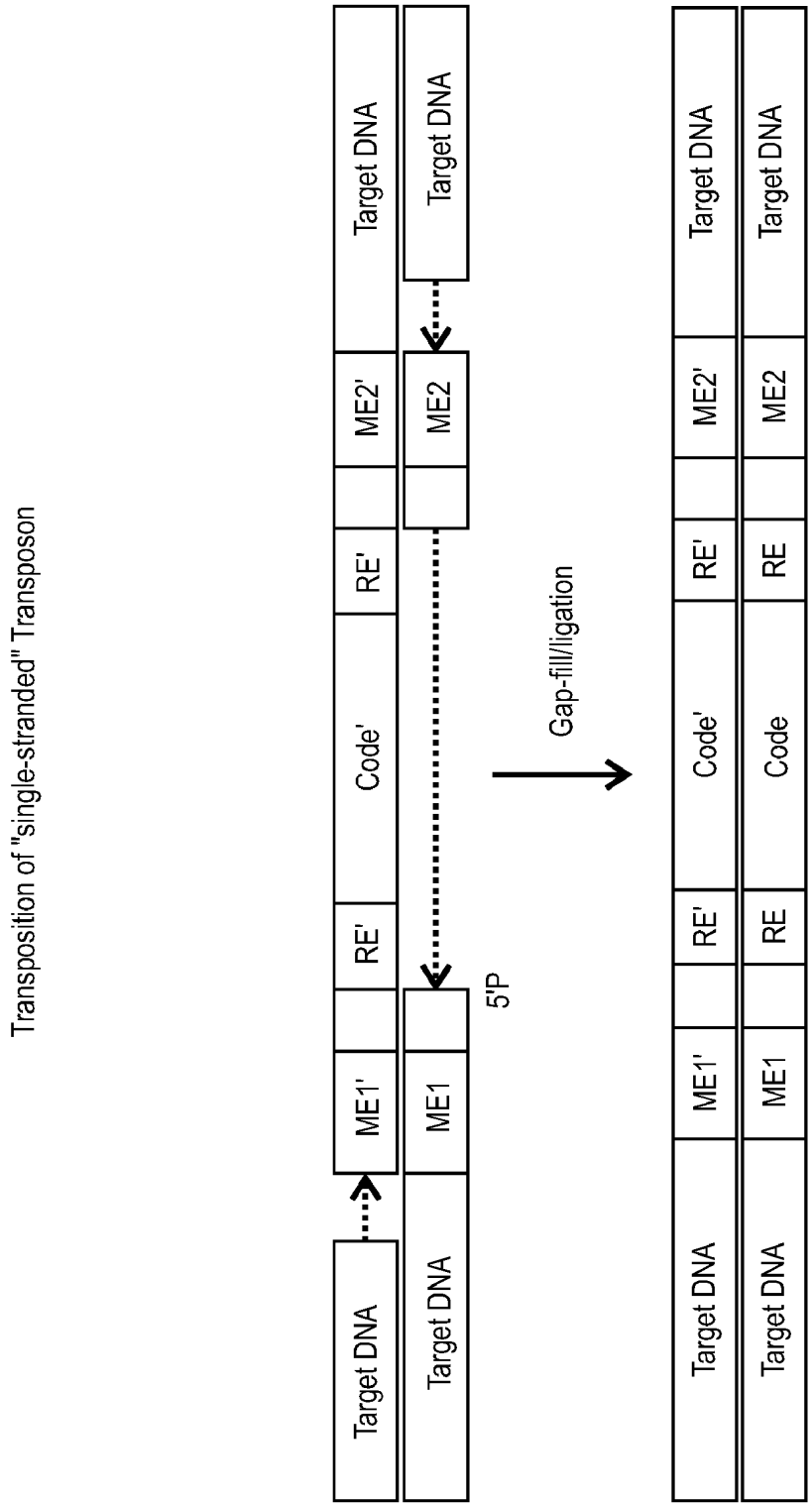


Fig. 6

IDT Exhibit 1018.0375

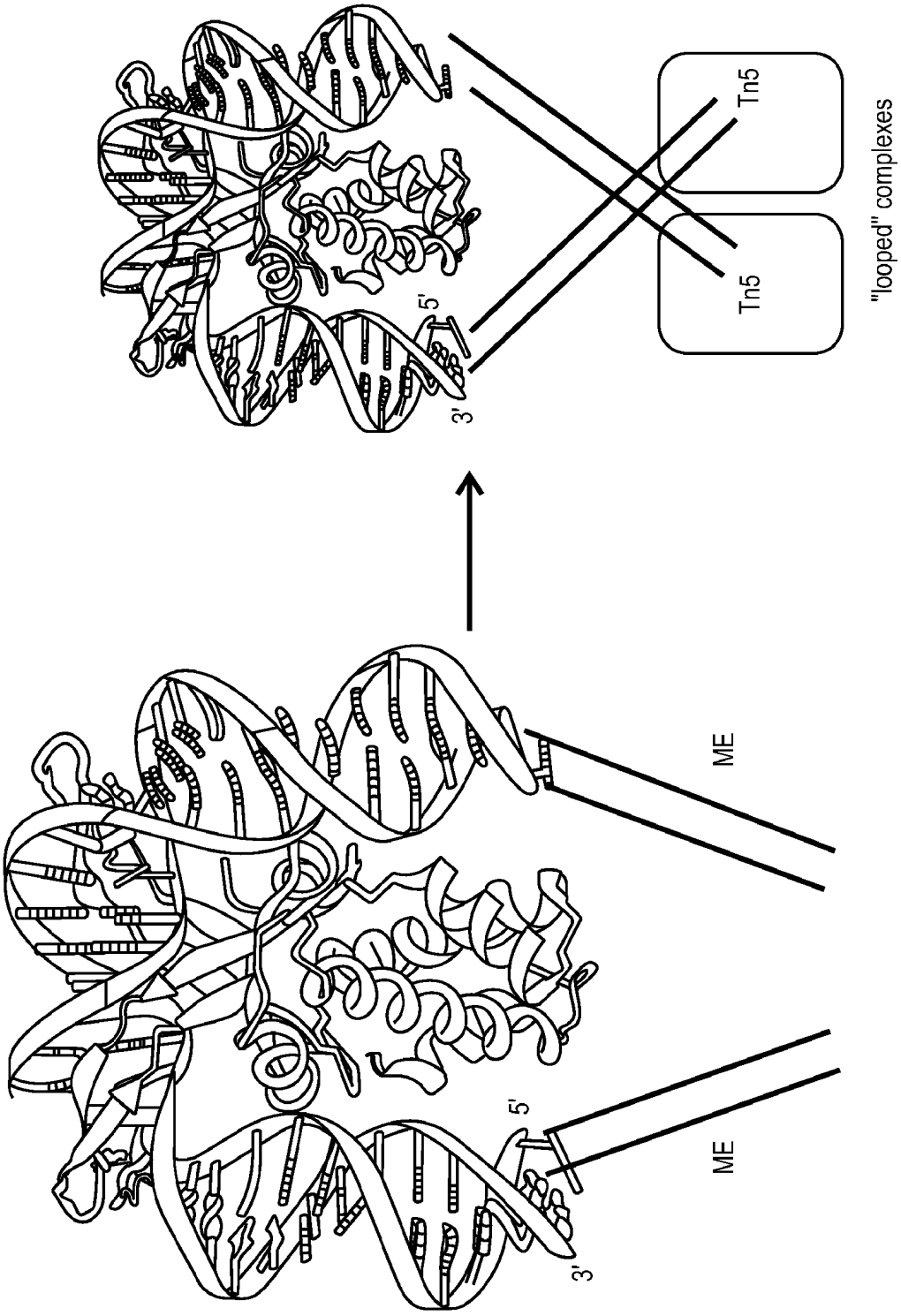


Fig. 7

IDT Exhibit 1018.0376

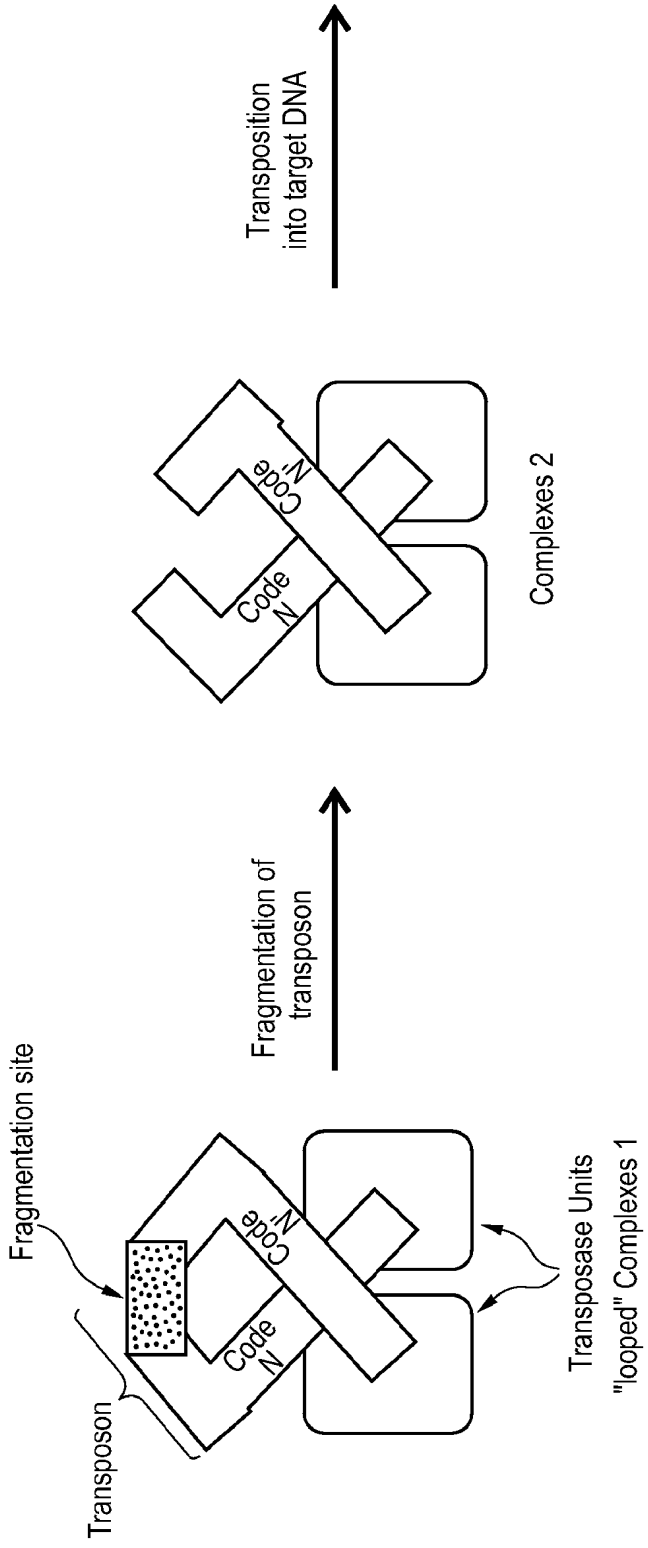


Fig. 8

IDT Exhibit 1018.0377

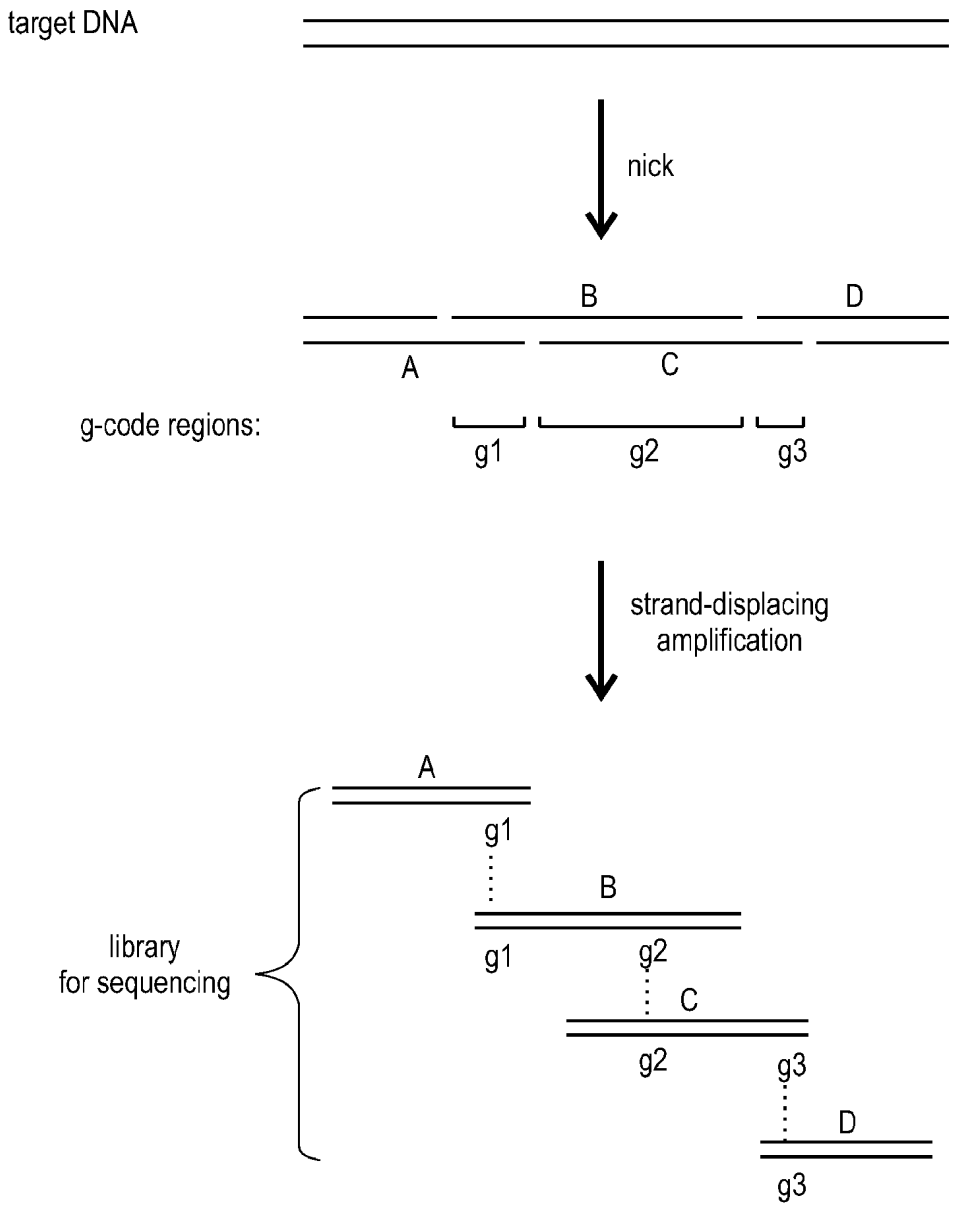


Fig. 9

IDT Exhibit 1018.0378

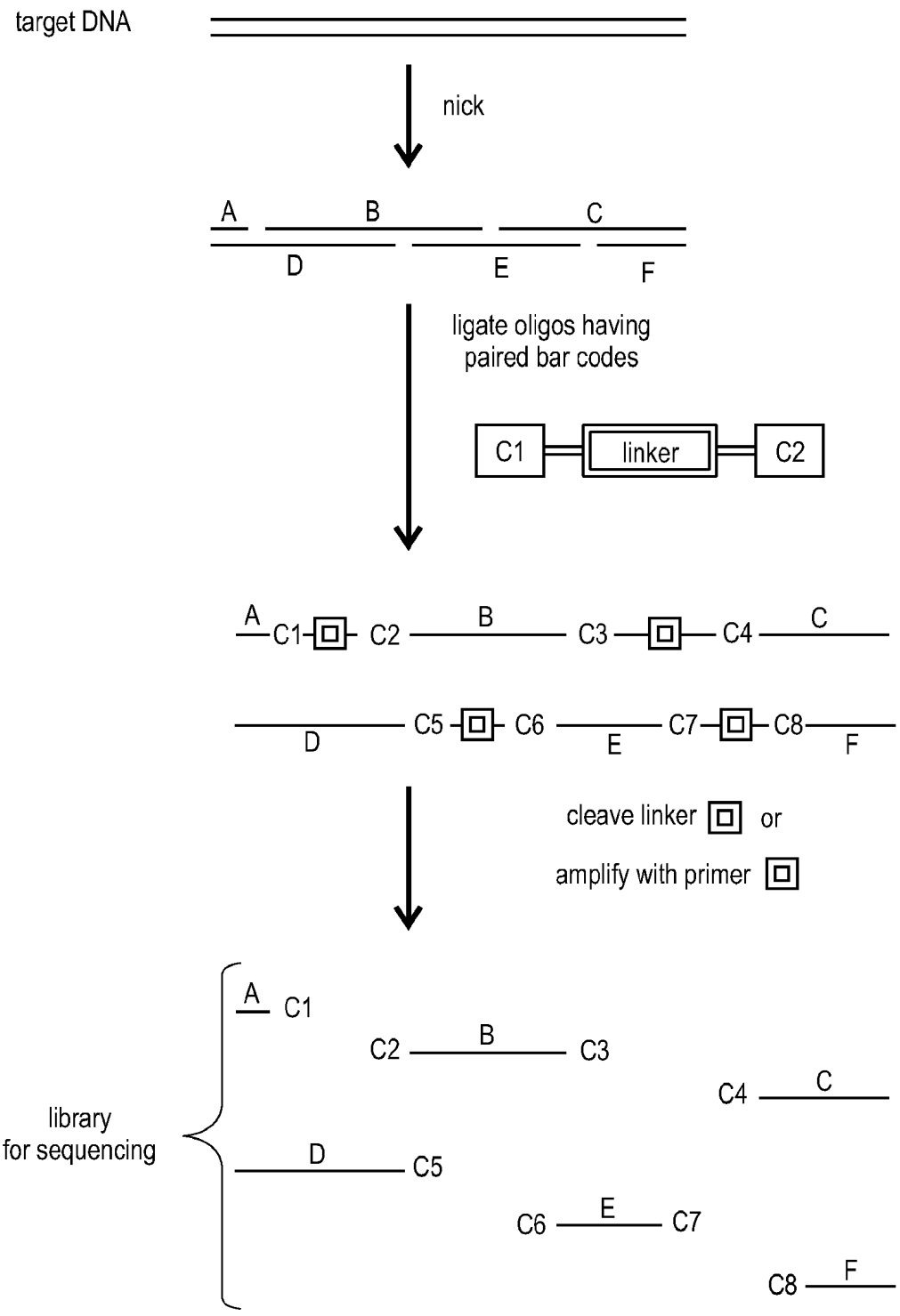
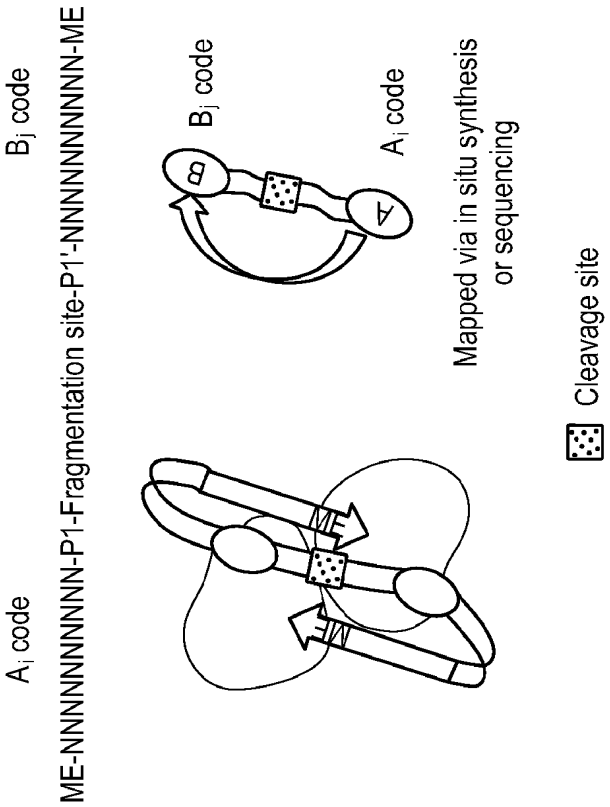


Fig. 10

IDT Exhibit 1018.0379



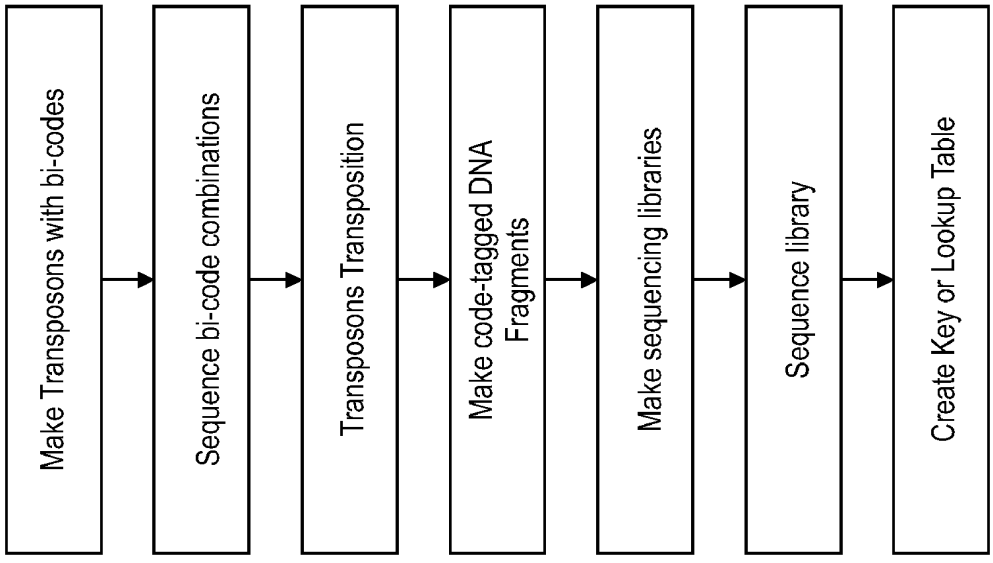


Fig. 11

IDT Exhibit 1018.0380

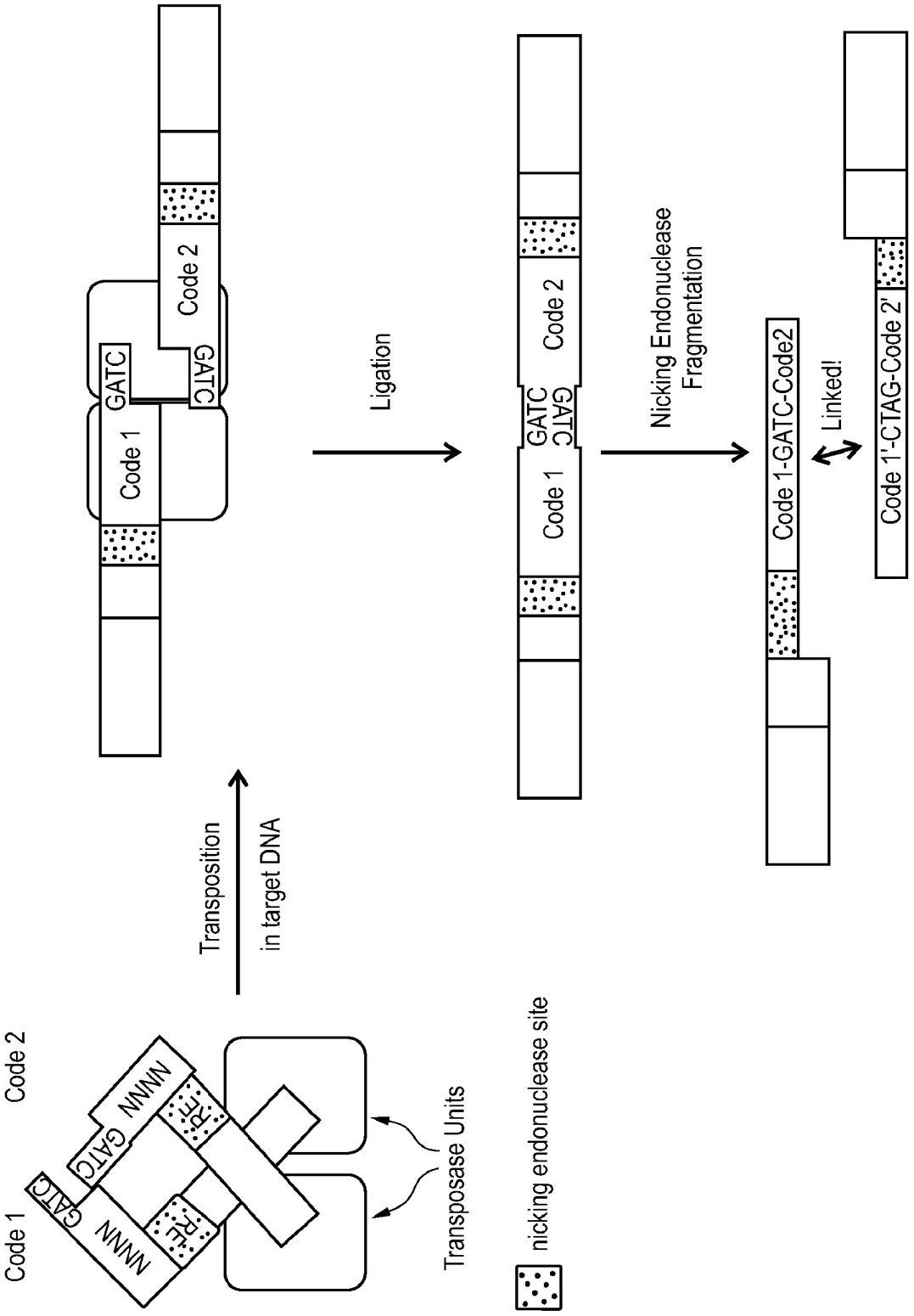

Code 2
NNNN
GATC
Code 1
NNNN
GATC
Transposase Units
nicking endonuclease site
Transposition
in target DNA
Code 1
GATC
Code 2
GATC
Ligation
Code 1
GATC
Code 2
GATC
Nicking Endonuclease
Fragmentation
Code 1-GATC-Code2
Linked!
Code 1'-CTAG-Code 2'

Fig. 12

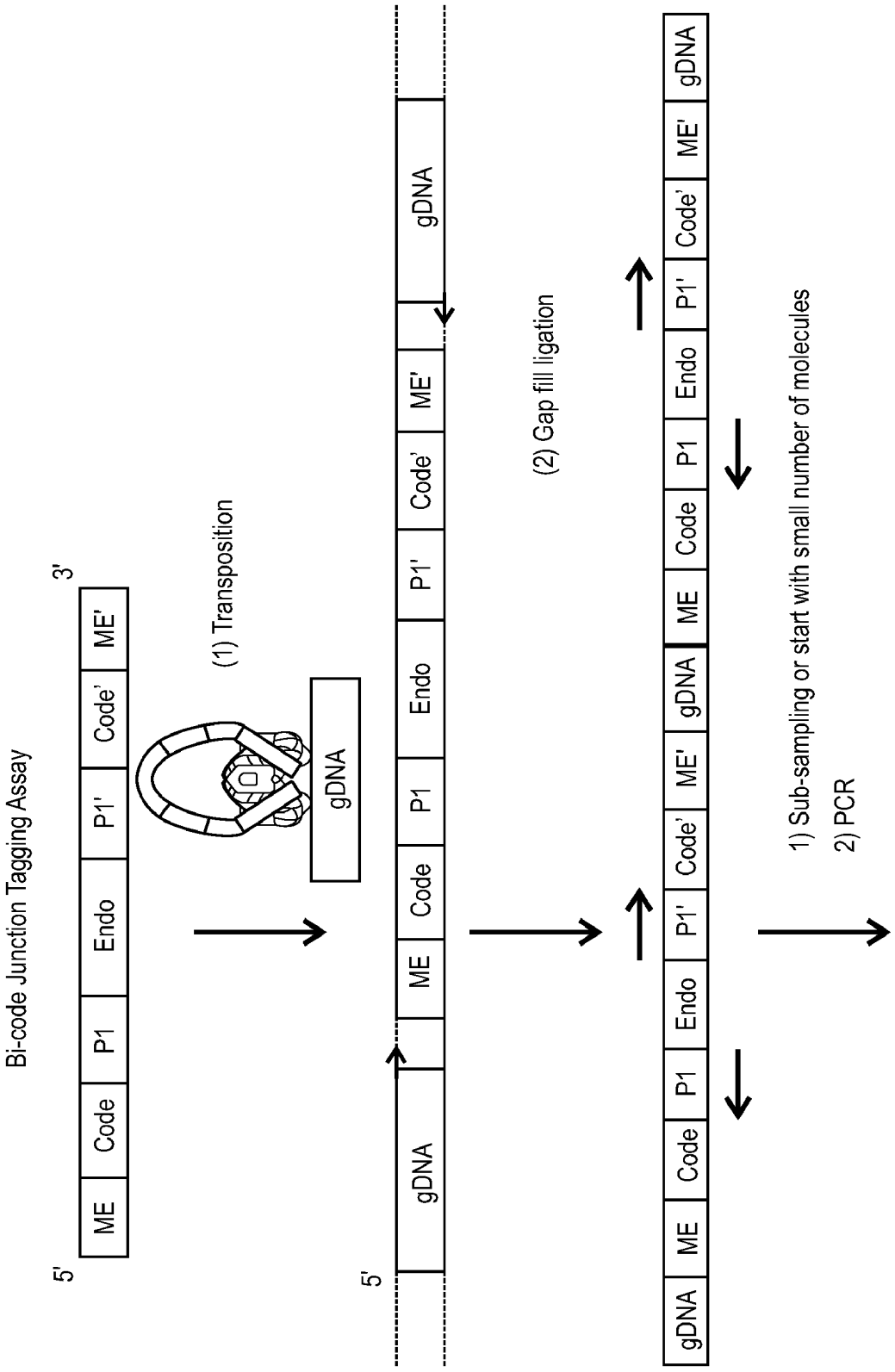

Bi-code Junction Tagging Assay
5'
3'
ME
Code
P1
Endo
P1'
Code'
ME'
(1) Transposition
gDNA
5'
gDNA
ME
Code
P1
Endo
P1'
Code'
ME'
gDNA
(2) Gap fill ligation
gDNA
ME
Code
P1
Endo
P1'
Code'
ME'
gDNA
ME
Code
P1
Endo
P1'
Code'
ME'
gDNA
1) Sub-sampling or start with small number of molecules
2) PCR

Fig. 13

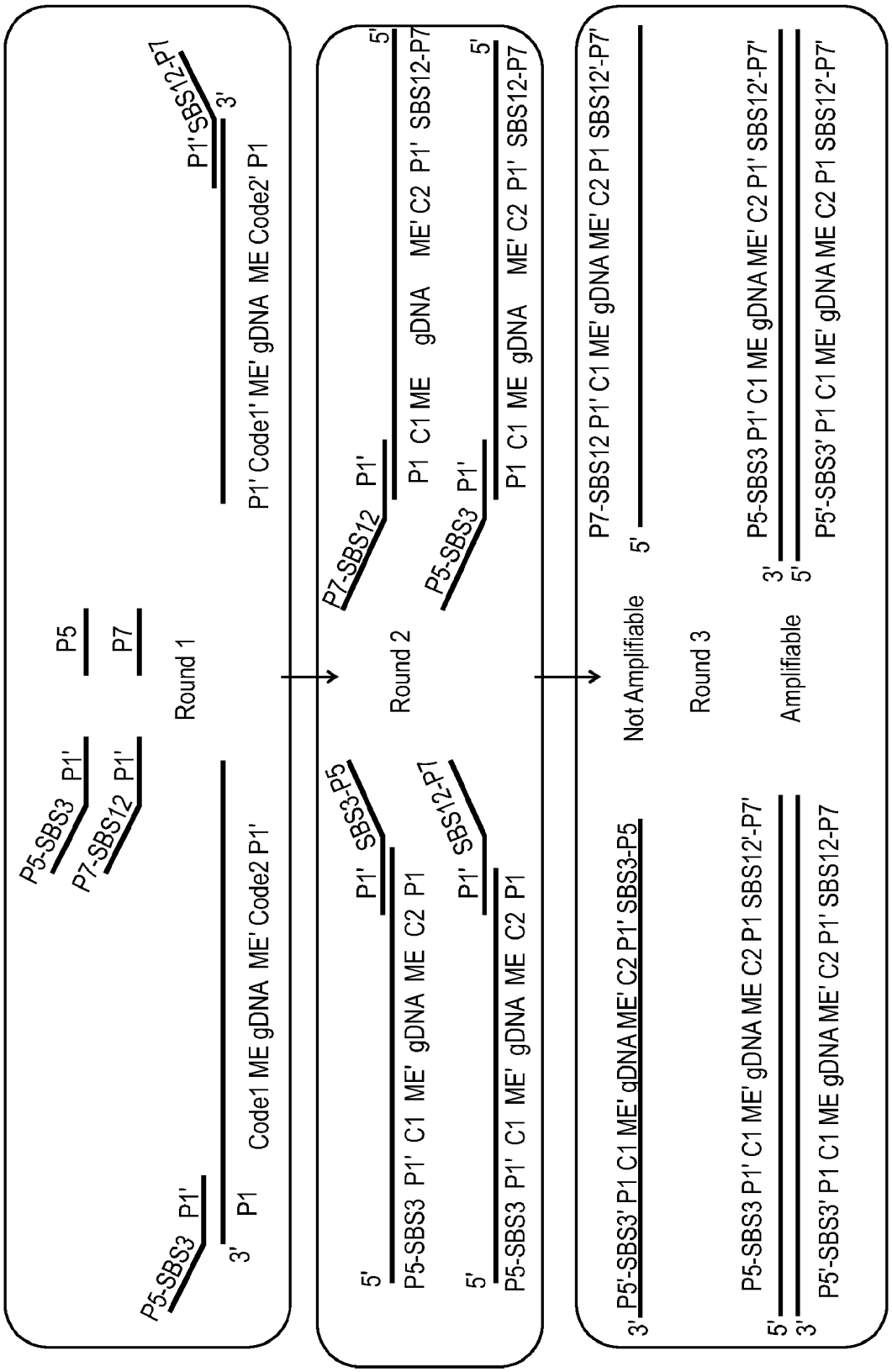

P5-SBS3 P1'
P7-SBS12 P1'
P5
P7
Round 1
P1'SBS12-P7
3'
P1' Code1' ME' gDNA ME Code2' P1
P5-SBS3 P1'
3' P1 Code1 ME gDNA ME' Code2 P1'
P7-SBS12 P1'
5'
P1 C1 ME gDNA ME' C2 P1' SBS12-P7
Round 2
P5-SBS3 P1'
5'
P1 C1 ME gDNA ME' C2 P1' SBS12-P7
P1' SBS3-P5
5'
P5-SBS3 P1' C1 ME' gDNA ME C2 P1
P1' SBS12-P7
5'
P5-SBS3 P1' C1 ME' gDNA ME C2 P1
P7-SBS12 P1' C1 ME' gDNA ME' C2 P1 SBS12'-P7'
5'
Not Amplifiable
3' P5'-SBS3' P1 C1 ME' gDNA ME' C2 P1' SBS3-P5
Round 3
P5-SBS3 P1' C1 ME' gDNA ME C2 P1 SBS12'-P7'
5'
3'
P5'-SBS3' P1 C1 ME gDNA ME' C2 P1' SBS12-P7
Amplifiable
P5-SBS3 P1' C1 ME gDNA ME' C2 P1' SBS12'-P7'
3'
5'
P5'-SBS3' P1 C1 ME' gDNA ME C2 P1 SBS12'-P7'

Fig. 14

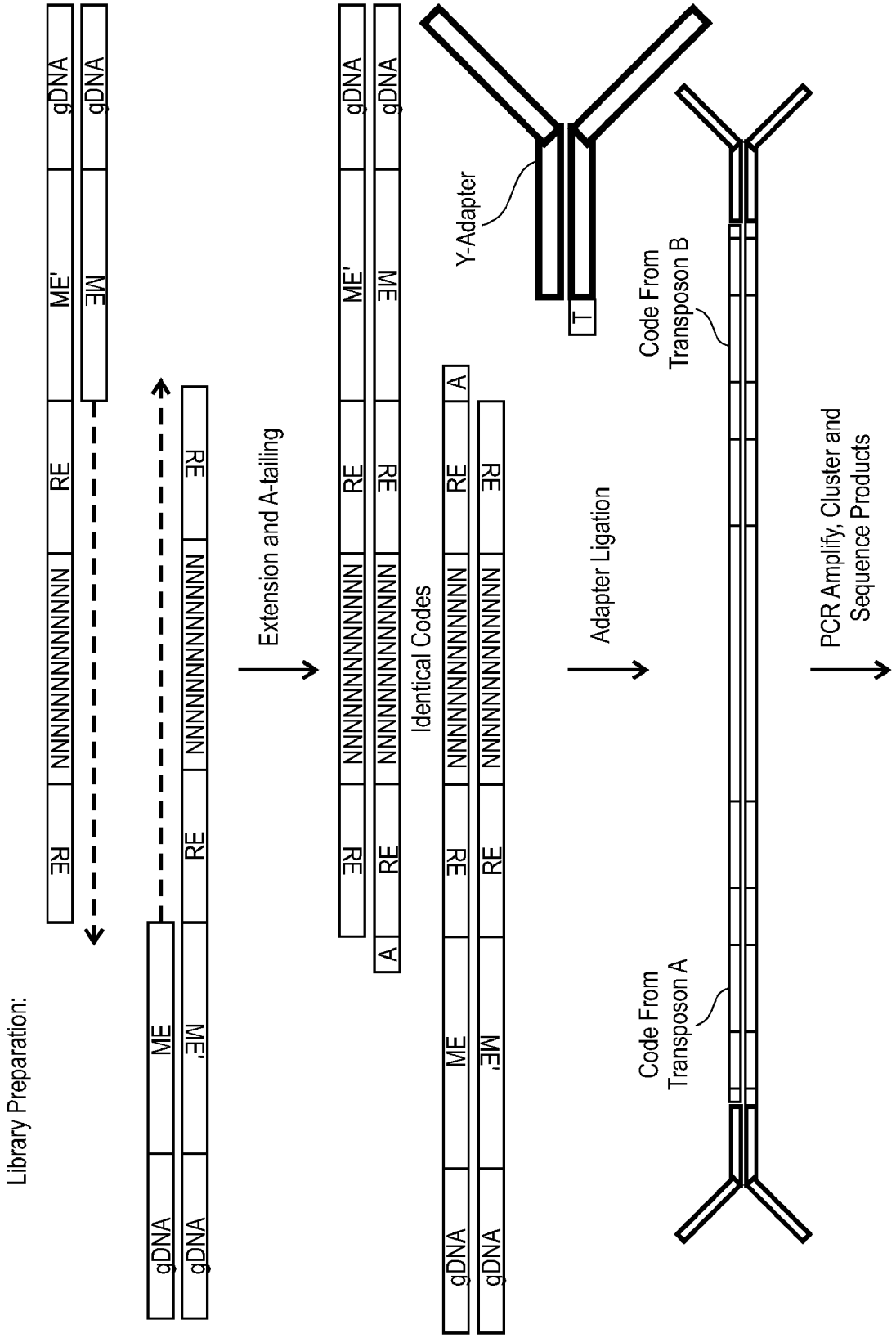
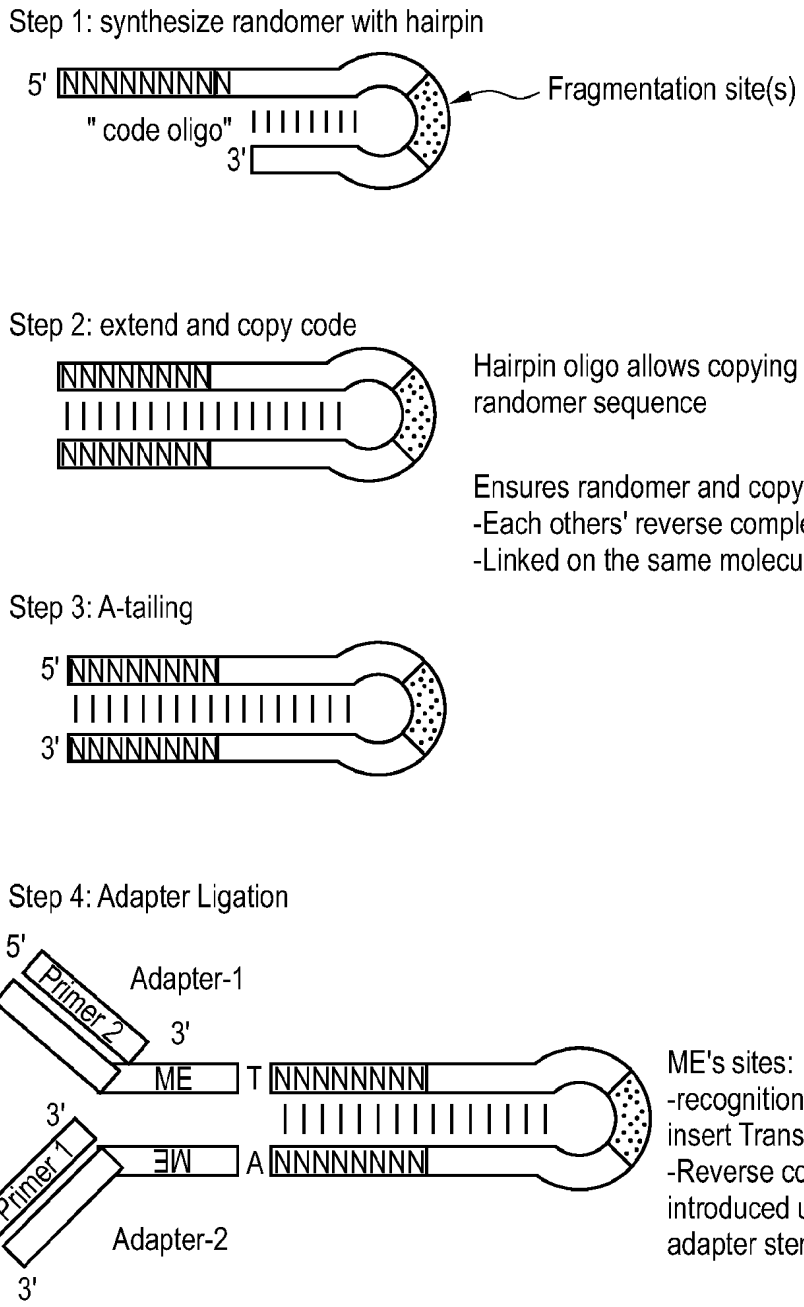

Library Preparation:
gDNA
ME
ME'
RE
NNNNNNNNNNNNNNN
Extension and A-tailing
A
Identical Codes
Y-Adapter
T
Adapter Ligation
Code From Transposon B
Code From Transposon A
PCR Amplify, Cluster and Sequence Products

Fig. 15



P2	ME	Code				Code'	ME	P1'
P2'	ME'	Code'				Code	ME'	P1

Fig. 16