



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

PUBLICATION NOTICE



OC000000075178206

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

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

Publication No.US-2015-0132763-A1

Publication Date:05/14/2015

NOTICE OF PUBLICATION OF APPLICATION

The above-identified application will be electronically published as a patent application publication pursuant to 37 CFR 1.211, et seq. The patent application publication number and publication date are set forth above.

The publication may be accessed through the USPTO's publicly available Searchable Databases via the Internet at www.uspto.gov. The direct link to access the publication is currently <http://www.uspto.gov/patft/>.

The publication process established by the Office does not provide for mailing a copy of the publication to applicant. A copy of the publication may be obtained from the Office upon payment of the appropriate fee set forth in 37 CFR 1.19(a)(1). Orders for copies of patent application publications are handled by the USPTO's Office of Public Records. The Office of Public Records can be reached by telephone at (703) 308-9726 or (800) 972-6382, by facsimile at (703) 305-8759, by mail addressed to the United States Patent and Trademark Office, Office of Public Records, Alexandria, VA 22313-1450 or via the Internet.

In addition, information on the status of the application, including the mailing date of Office actions and the dates of receipt of correspondence filed in the Office, may also be accessed via the Internet through the Patent Electronic Business Center at www.uspto.gov using the public side of the Patent Application Information and Retrieval (PAIR) system. The direct link to access this status information is currently <http://pair.uspto.gov/>. Prior to publication, such status information is confidential and may only be obtained by applicant using the private side of PAIR.

Further assistance in electronically accessing the publication, or about PAIR, is available by calling the Patent Electronic Business Center at 1-866-217-9197.

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



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	05/20/2015		
COOLEY LLP ATTN: Patent Group 1299 Pennsylvania Avenue, NW Suite 700 Washington, DC 20004			EXAMINER STRZELECKA, TERESA E	
			ART UNIT 1637	PAPER NUMBER
			NOTIFICATION DATE 05/20/2015	DELIVERY MODE 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):

zpatcdocketing@cooley.com

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. Applicant's election with traverse of Group I (claims 1-20; species A, E and G) in the reply filed on March 6, 2015 is acknowledged. The traversal is on the ground(s) that:

“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.”

Applicant did not argue the election of species.

Since the art used for rejection of claims of Group I teaches the kit of Group II, both groups are examined together.

3. Claims 4-6 and 8 are withdrawn from further consideration pursuant to 37 CFR 1.142(b), as being drawn to a nonelected species, there being no allowable generic or linking claim. Applicant timely traversed the restriction (election) requirement in the reply filed on March 6, 2015.
4. Claims 1-3, 7 and 9-21 will be examined.

Information Disclosure Statement

5. The information disclosure statement (IDS) submitted on December 4, 2014 is in compliance with the provisions of 37 CFR 1.97. Accordingly, the information disclosure statement is being considered by the examiner.

6. The information disclosure statement (IDS) submitted on March 31, 2015 was filed after the mailing date of the Restriction/Election requirement on January 7, 2015. The submission is in compliance with the provisions of 37 CFR 1.97. Accordingly, the information disclosure statement is being considered by the examiner.

Claim Interpretation

7. Applicant defined the term “adaptor” in paragraph [0027] as follows:

“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.”

Therefore, since any oligonucleotide can perform such function, any oligonucleotide is considered to anticipate this term.

8. Applicant defined the term “identifier site” in paragraph [0030] as follows:

“An identifier site is a nucleic acid sequence that comprises random bases and is used to identify duplicate sequencing reads.”

9. Applicant did not define the term “indexing site”, therefore it is considered to be any sequence.

10. Applicant defined the term “universal primer” in paragraph [0035] as follows:

“A universal primer is a primer that can be used for amplifying a number of different sequences.”

11. There is no requirement in the claims for the indexing primer binding site, indexing site, identifier site and target sequence primer binding site to have different sequences, in other words, target sequence binding site could be the identifier site.

12. Applicant did not define the term “kit”, therefore any plurality of claimed oligonucleotides is considered to anticipate this term.

Claim Rejections - 35 USC § 101

13. 35 U.S.C. 101 reads as follows:

Whoever invents or discovers any new and useful process, machine, manufacture, or composition of matter, or any new and useful improvement thereof, may obtain a patent therefor, subject to the conditions and requirements of this title.

14. Claim 21 is rejected under 35 U.S.C. 101 because the claimed invention is directed to a judicial exception (i.e., a law of nature, a natural phenomenon, or an abstract idea) without significantly more. Claim(s) 21 is directed to a plurality of adaptors, i.e., nucleic acid sequences, comprising an indexing primer binding site, an indexing site, an identifier site and a target sequencing binding site. No specific sequences are claimed; therefore the claim is drawn to a plurality of oligonucleotides which have corresponding sequences occurring in nature. The

claim(s) does/do not include additional elements that are sufficient to amount to significantly more than the judicial exception because no other structural elements are claimed.

In conclusion, claim 21 is not patent eligible under 35 U.S.C. 101.

Claim Rejections - 35 USC § 102

15. In the event the determination of the status of the application as subject to AIA 35 U.S.C. 102 and 103 (or as subject to pre-AIA 35 U.S.C. 102 and 103) is incorrect, any correction of the statutory basis for the rejection will not be considered a new ground of rejection if the prior art relied upon, and the rationale supporting the rejection, would be the same under either status.

16. The following is a quotation of the appropriate paragraphs of 35 U.S.C. 102 that form the basis for the rejections under this section made in this Office action:

A person shall be entitled to a patent unless –

(a)(1) the claimed invention was patented, described in a printed publication, or in public use, on sale or otherwise available to the public before the effective filing date of the claimed invention.

17. Claim 21 is rejected under 35 U.S.C. 102(a)(1) as being anticipated by Grothues et al. (Nucl. Acids Res., vol. 21, pp. 1321-1322, 1993).

Regarding claim 21, Grothues et al. teach a plurality of tagged random primers (= adaptors) containing 9-15 bp random 3' part and a 17 bp constant sequence at the 5' end (page 1321, second paragraph). Therefore, the primers of Grothues et al. contain an indexing primer binding site, an indexing site (both contained in the 5' constant sequence), and an identifier site with random bases which is also a target binding site.

18. Claim 21 is rejected under 35 U.S.C. 102(a)(1) as being anticipated by McCloskey et al. (Biochem. Genet., vol. 45, pp. 761-767, 2007).

Regarding claim 21, McCloskey et al. teach plurality of randomly-barcoded oligonucleotides (= adaptors) containing, from the 3' end, a target-specific sequence, batch-stamp sequence, random 7 bp barcode (= identifier), another primer binding site and another sequence (Fig. 1). Therefore, the oligonucleotides of McCloskey et al. contain an indexing primer binding site, an indexing site, an identifier site with random bases which is also a target binding site.

Claim Rejections - 35 USC § 103

19. In the event the determination of the status of the application as subject to AIA 35 U.S.C. 102 and 103 (or as subject to pre-AIA 35 U.S.C. 102 and 103) is incorrect, any correction of the statutory basis for the rejection will not be considered a new ground of rejection if the prior art relied upon, and the rationale supporting the rejection, would be the same under either status.

20. The following is a quotation of 35 U.S.C. 103 which forms the basis for all obviousness rejections set forth in this Office action:

A patent for a claimed invention may not be obtained, notwithstanding that the claimed invention is not identically disclosed as set forth in section 102 of this title, if the differences between the claimed invention and the prior art are such that the claimed invention as a whole would have been obvious before the effective filing date of the claimed invention to a person having ordinary skill in the art to which the claimed invention pertains. Patentability shall not be negated by the manner in which the invention was made.

21. Claims 1-3, 7 and 9-20 are rejected under 35 U.S.C. 103 as being unpatentable over Samuels et al. (US 2012/0220494 A1; filed February 16, 2012 and published August 30, 2012) and McCloskey et al. (Biochem. Genet., vol. 45, pp. 761-767, 2007).

A) Regarding claims 1 and 7, Samuels et al. teach 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 (page 14, [0205]; page 15, [0209], [0212], [0216]-[0217]; Fig. 6; page 16, [0231]-[0232]; page 17, [0232]-[0235]; page 16/17, [0245]; where the adaptors have an indexing primer binding site, an indexing site, an identifier site and a target sequence primer binding site);

b) amplifying the adapter-nucleic acid fragment ligated products (page 22, [0288]-[0289]; page 25, [0326]-[0327]);

c) generating a population of sequencing reads from amplified adapter-nucleic acid fragment ligated products (page 25, [0326]-[0328]; page 26, [0328]-[0337]; and

d) detecting the population of sequencing reads comprising a sequencing read with a duplicate identifier site and target sequence (page 26/27, [0328]; page 27, [0339]-[0341]; Fig. 43 and 44; where sequencing errors due to PCR are detected).

Regarding claim 3, Samuels et al. teach sequencing the identifier site with the indexing site (Fig. 43; page 27, [0339]-[0341]).

Regarding claim 9, Samuels et al. teach barcoding multiple samples (page 6, [0122]; page 14, [0202]).

Regarding claims 13 and 14, Samuels et al. teach DNA and RNA fragments, and genomic DNA (page 2, [0017]).

Regarding claims 17 and 18, Samuels et al. teach identifier site between 1 and 8 nucleotides in length (page 3, [0029]; page 15/16, [0217]).

Regarding claim 19, Samuels et al. teach indexing primer site being a universal primer site (page 15, [0216]).

Regarding claim 20, Samuels et al. teach target sequence binding site being a universal binding site (page 15, [0216]).

B) Samuels et al. teach detection of PCR-introduced errors in sequencing, but do not specifically teach detecting duplicate sequencing reads.

C) Regarding claim 1, McCloskey et al. teach attaching an adaptor containing an indexing primer binding site, an indexing site, an identifier site and a target sequence binding site to genomic DNA by amplification, PCR-amplifying the resulting construct and sequencing (Fig. 1; page 762, paragraphs 3-5; page 763; page 764, paragraphs 1-4).

Regarding claims 1, 2 and 10-12, McCloskey et al. teach sequencing the PCR products containing the adapter sequences, grouping the sequencing reads based on the batch stamp and barcode and detecting and removing duplicate reads (page 766, fourth and fifth paragraphs).

Regarding claims 15 and 16, McCloskey et al. teach indexing site between six and eight bp (page 764, second paragraph).

It would have been *prima facie* obvious to one of ordinary skill in the art at the time of the invention to have used the batch-stamp and barcode-encoded oligonucleotides as suggested by McCloskey et al. in the method of Samuels et al. to eliminate PCR-originating artefacts. As stated by McCloskey et al. (Abstract):

“Polymerase chain reaction (PCR) has become the mainstay of DNA sequence analysis. Yet there is always uncertainty concerning the source of the template DNA that gave rise to a particular PCR product. The risks of contamination, biased amplification, and product redundancy are especially high when limited amounts of template DNA are used. We have developed and applied molecular encoding principles to solve this source-uncertainty problem for DNA sequences generated by standard PCR. Batch-stamps specify the date and sample identity, and barcodes detect template redundancy. Our approach thus enables classification of each PCR-derived sequence as valid, contaminant, or redundant, and provides a measure of sequence diversity. We recommend that batch-stamps and barcodes be used when amplifying irreplaceable DNAs and cDNAs available for forensic, clinical, single cell, and ancient DNA analyses.”

22. No claims are allowed.

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
May 15, 2015

Search Notes 	Application/Control No. 14540917	Applicant(s)/Patent Under Reexamination AMORESE ET AL.
	Examiner TERESA E STRZELECKA	Art Unit 1637

CPC- SEARCHED			
Symbol		Date	Examiner
C12Q1/6869 C12Q2563/179 C12Q2535/122 C12Q2537/1376 C12Q2563/159 C12Q1/6806 C12Q2521/319 C12Q2521/507 C12Q2525/15 C12Q2525/155 C12Q2525/161 C12Q2525/185 C12Q2525/191 C12Q2525/197		5/13/15	/TS/

CPC COMBINATION SETS - SEARCHED		
Symbol	Date	Examiner

US CLASSIFICATION SEARCHED			
Class	Subclass	Date	Examiner

SEARCH NOTES		
Search Notes	Date	Examiner
IDS (12/4/14; 3/21/15)	5/13/15	/TS/
WEST	5/13/15	/TS/
Google	5/13/15	/TS/
Inventor name search	5/13/15	/TS/

INTERFERENCE SEARCH			
US Class/ CPC Symbol	US Subclass / CPC Group	Date	Examiner

--	--

<i>Index of Claims</i> 	Application/Control No. 14540917	Applicant(s)/Patent Under Reexamination AMORESE ET AL.
	Examiner TERESA E STRZELECKA	Art Unit 1637

✓	Rejected	-	Cancelled	N	Non-Elected	A	Appeal
=	Allowed	÷	Restricted	I	Interference	O	Objected

☐ Claims renumbered in the same order as presented by applicant			☐ CPA			☐ T.D.			☐ R.1.47		
CLAIM		DATE									
Final	Original	05/14/2015									
	1	✓									
	2	✓									
	3	✓									
	4	N									
	5	N									
	6	N									
	7	✓									
	8	N									
	9	✓									
	10	✓									
	11	✓									
	12	✓									
	13	✓									
	14	✓									
	15	✓									
	16	✓									
	17	✓									
	18	✓									
	19	✓									
	20	✓									
	21	✓									

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 Teresa Strzelecka	
	Attorney Docket Number	NUGN-001/01US	

ALL REFERENCES CONSIDERED EXCEPT WHERE LINED THROUGH. /TS/

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	14540917 - GAU: 1637
Filing Date	2014-11-13	
First Named Inventor	Douglas AMORESE	
Art Unit	1636	
Examiner Name	Unassigned	Teresa
Attorney Docket Number	NUGN-001/01US Strzelecka	

ALL REFERENCES CONSIDERED EXCEPT WHERE LINED THROUGH. /TS/

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	14540917 - GAU: 1637
	Filing Date		2014-11-13	
	First Named Inventor	Douglas AMORESE		
	Art Unit	4000		
	Examiner Name	Unassigned	Teresa	
	Attorney Docket Number	NUGN-001/01US	Strzelecka	

ALL REFERENCES CONSIDERED EXCEPT WHERE LINED THROUGH. /TS/

	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 Code ¹	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	14540917 - GAU: 1637
Filing Date	2014-11-13	
First Named Inventor	Douglas AMORESE	
Art Unit	1006	
Examiner Name	Unassigned	Teresa
Attorney Docket Number	NUGN-001/01US	Strzelecka

ALL REFERENCES CONSIDERED EXCEPT WHERE LINED THROUGH. /TS/

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	14540917 - GAU: 1637
	Filing Date		2014-11-13	
	First Named Inventor	Douglas AMORESE		
	Art Unit	1636		
	Examiner Name	Unassigned	Teresa	
	Attorney Docket Number	NUGN-001/01US	Strzelecka	

ALL REFERENCES CONSIDERED EXCEPT WHERE LINED THROUGH. /TS/

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	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
								☐

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.	☐

INFORMATION DISCLOSURE STATEMENT BY APPLICANT (Not for submission under 37 CFR 1.99)	Application Number		14540917	14540917 - GAU: 1637
	Filing Date		2014-11-13	
	First Named Inventor	Douglas AMORESE		
	Art Unit	4636		
	Examiner Name	Unassigned Teresa Strzelecka		
	Attorney Docket Number	NUGN-001/01US Strzelecka		

ALL REFERENCES CONSIDERED EXCEPT WHERE LINED THROUGH. /TS/

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	/Teresa Strzelecka/	Date Considered	05/13/2015
--------------------	---------------------	-----------------	------------

*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.

Notice of References Cited	Application/Control No. 14/540,917		Applicant(s)/Patent Under Reexamination AMORESE ET AL.	
	Examiner TERESA E. STRZELECKA		Art Unit 1637	Page 1 of 1

U.S. PATENT DOCUMENTS

*		Document Number Country Code-Number-Kind Code	Date MM-YYYY	Name	Classification
*	A	US-2012/0220494	08-2012	Samuels et al.	506/16
	B	US-			
	C	US-			
	D	US-			
	E	US-			
	F	US-			
	G	US-			
	H	US-			
	I	US-			
	J	US-			
	K	US-			
	L	US-			
	M	US-			

FOREIGN PATENT DOCUMENTS

*		Document Number Country Code-Number-Kind Code	Date MM-YYYY	Country	Name	Classification
	N					
	O					
	P					
	Q					
	R					
	S					
	T					

NON-PATENT DOCUMENTS

*		Include as applicable: Author, Title Date, Publisher, Edition or Volume, Pertinent Pages)
	U	Grothues, D. et al., Nucl. Acids Res., vol. 21, pp. 1321-1322 (1993)
	V	McCloskey, M.L. et al., Biochem. Genet., vol. 45, pp. 761-767 (2007)
	W	
	X	

*A copy of this reference is not being furnished with this Office action. (See MPEP § 707.05(a).)
Dates in MM-YYYY format are publication dates. Classifications may be US or foreign.



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

BIB DATA SHEET

CONFIRMATION NO. 7076

SERIAL NUMBER	FILING or 371(c) DATE	CLASS	GROUP ART UNIT	ATTORNEY DOCKET NO.		
14/540,917	11/13/2014	435	1637	NUGN-001/01US		
RULE						
APPLICANTS Nugen Technologies, Inc., San Carlos, CA INVENTORS Douglas AMORESE, Los Altos, CA; Jonathan Scolnick, San Francisco, CA; Ben Schroeder, San Mateo, CA; ** CONTINUING DATA ***** This appln claims benefit of 61/903,826 11/13/2013 ** FOREIGN APPLICATIONS ***** ** IF REQUIRED, FOREIGN FILING LICENSE GRANTED ** 11/25/2014						
Foreign Priority claimed ☐ Yes ☒ No	35 USC 119(a-d) conditions met ☐ Yes ☐ No	☐ Met after Allowance	STATE OR COUNTRY	SHEETS DRAWINGS	TOTAL CLAIMS	INDEPENDENT CLAIMS
Verified and /TERESA E STRZELECKA/	Examiner's Signature	Initials	CA	6	21	2
ADDRESS COOLEY LLP ATTN: Patent Group 1299 Pennsylvania Avenue, NW Suite 700 Washington, DC 20004 UNITED STATES						
TITLE COMPOSITIONS AND METHODS FOR IDENTIFICATION OF A DUPLICATE SEQUENCING READ						
FILING FEE RECEIVED 1820	FEES: Authority has been given in Paper No. _____ to charge/credit DEPOSIT ACCOUNT No. _____ for following:		☐ All Fees			
			☐ 1.16 Fees (Filing)			
			☐ 1.17 Fees (Processing Ext. of time)			
			☐ 1.18 Fees (Issue)			
			☐ Other _____			
		☐ Credit				

WEST Search History for Application 14540917

Creation Date: 2015051314:35

Prior Art Searches

Query	DB	Op.	Plur.	Thes.	Date
sequencing and reads near3 (duplicat\$7 or redundan\$8) and (barcod\$8 or index\$8) near3 random and (duplicat\$7 or redundan\$8) near3 remov\$8	PGPB	ADJ	YES	ASSIGNEE	05-13-2015
sequencing and reads near3 (duplicat\$7 or redundan\$8) and (barcod\$8 or index\$8) near3 (random\$8 or degener\$8)	PGPB	ADJ	YES	ASSIGNEE	05-13-2015
sequencing and reads near3 (duplicat\$7 or redundan\$8) and (barcod\$8 or index\$8) near3 (random\$8 or degener\$8)	USPT	ADJ	YES	ASSIGNEE	05-13-2015
sequencing and duplicate near3 remov\$8 and ((C12Q1/6869 C12Q2563/179 C12Q2535/122 C12Q2537/1376 C12Q2563/159 C12Q1/6806 C12Q2521/319 C12Q2521/507 C12Q2525/15 C12Q2525/155 C12Q2525/161 C12Q2525/185 C12Q2525/191 C12Q2525/197 C12Q2525/313 C12Q2537/159 C12Q2537/165 C12Q2549/10 C12Q2565/514 C12Q1/6874 C12Q1/6876 C12Q1/6879 C12Q1/689 C12Q2600/156 C12Q2600/166 C12N15/1065 C12N15/1093 C12N15/1089 G06F19/18).CPC.)	PGPB	ADJ	YES	ASSIGNEE	05-13-2015
(sequencing and duplicate near3 remov\$8 and ((C12Q1/6869 C12Q2563/179 C12Q2535/122 C12Q2537/1376 C12Q2563/159 C12Q1/6806 C12Q2521/319 C12Q2521/507 C12Q2525/15 C12Q2525/155 C12Q2525/161 C12Q2525/185 C12Q2525/191 C12Q2525/197 C12Q2525/313 C12Q2537/159 C12Q2537/165 C12Q2549/10 C12Q2565/514 C12Q1/6874 C12Q1/6876 C12Q1/6879 C12Q1/689 C12Q2600/156 C12Q2600/166 C12N15/1065 C12N15/1093 C12N15/1089 G06F19/18).CPC.)) and random\$8 near3 (barcod\$8 or index\$8)	PGPB	ADJ	YES	ASSIGNEE	05-13-2015
sequencing and duplicate near3 remov\$8 and ((C12Q1/6869 C12Q2563/179 C12Q2535/122 C12Q2537/1376 C12Q2563/159 C12Q1/6806 C12Q2521/319 C12Q2521/507 C12Q2525/15 C12Q2525/155 C12Q2525/161 C12Q2525/185	USPT	ADJ	YES	ASSIGNEE	05-13-2015

C12Q2525/191 C12Q2525/197 C12Q2525/313 C12Q2537/159 C12Q2537/165 C12Q2549/10 C12Q2565/514 C12Q1/6874 C12Q1/6876 C12Q1/6879 C12Q1/689 C12Q2600/156 C12Q2600/166 C12N15/1065 C12N15/1093 C12N15/1089 G06F19/18).CPC.)					
(sequencing and duplicate near3 remov\$8 and ((C12Q1/6869 C12Q2563/179 C12Q2535/122 C12Q2537/1376 C12Q2563/159 C12Q1/6806 C12Q2521/319 C12Q2521/507 C12Q2525/15 C12Q2525/155 C12Q2525/161 C12Q2525/185 C12Q2525/191 C12Q2525/197 C12Q2525/313 C12Q2537/159 C12Q2537/165 C12Q2549/10 C12Q2565/514 C12Q1/6874 C12Q1/6876 C12Q1/6879 C12Q1/689 C12Q2600/156 C12Q2600/166 C12N15/1065 C12N15/1093 C12N15/1089 G06F19/18).CPC.)) and random\$8 near3 (barcod\$8 or index\$8)	USPT	ADJ	YES	ASSIGNEE	05-13-2015
sequencing and duplicate near3 remov\$8 and ((C12Q1/6869 C12Q2563/179 C12Q2535/122 C12Q2537/1376 C12Q2563/159 C12Q1/6806 C12Q2521/319 C12Q2521/507 C12Q2525/15 C12Q2525/155 C12Q2525/161 C12Q2525/185 C12Q2525/191 C12Q2525/197 C12Q2525/313 C12Q2537/159 C12Q2537/165 C12Q2549/10 C12Q2565/514 C12Q1/6874 C12Q1/6876 C12Q1/6879 C12Q1/689 C12Q2600/156 C12Q2600/166 C12N15/1065 C12N15/1093 C12N15/1089 G06F19/18).CPC.)	DWPI	ADJ	YES	ASSIGNEE	05-13-2015
20110269647	PGPB	ADJ	YES	ASSIGNEE	05-13-2015
(Amorese.in. or Scolnick.in. or Schroeder.in. or Nugen.asn.) and sequencing and duplicate	PGPB	ADJ	YES	ASSIGNEE	05-13-2015
(Amorese.in. or Scolnick.in. or Schroeder.in. or Nugen.asn.) and sequencing and duplicate	USPT	ADJ	YES	ASSIGNEE	05-13-2015
US-6251639-\$.DID. OR US-5750341-\$.DID. OR US-6306597-\$.DID. OR US-5969119-\$.DID. OR US-7244559-\$.DID. OR US-7335762-\$.DID. OR US-7211390-\$.DID. OR US-7244567-\$.DID.	USPT	ADJ	YES	ASSIGNEE	05-13-2015
US-7264929-\$.DID. OR US-7323305-\$.DID. OR US-7501245-\$.DID. OR US-7491498-\$.DID. OR US-7276720-\$.DID. OR US-7462452-\$.DID. OR US-7476504-\$.DID. OR US-7405281-\$.DID. OR US-7170050-\$.DID. OR US-7462468-\$.DID. OR US-7476503-\$.DID.	USPT	ADJ	YES	ASSIGNEE	05-13-2015
	USPT	ADJ	YES	ASSIGNEE	05-13-2015

US-7315019-\$.DID. OR US-7302146-\$.DID. OR US-7313318-\$.DID. OR US-6449562-\$.DID. OR US-6287766-\$.DID. OR US-7361468-\$.DID. OR US-6225109-\$.DID. OR US-6110709-\$.DID.					
US-20130231253-\$.DID. OR US-20060024711-\$.DID. OR US-20090061439-\$.DID. OR US-20080087826-\$.DID. OR US-20060286566-\$.DID. OR US-20060024678-\$.DID. OR US-20080213770-\$.DID. OR US-20080103058-\$.DID. OR US-20090029385-\$.DID. OR US-20090068655-\$.DID.	PGPB	ADJ	YES	ASSIGNEE	05-13-2015
US-20090024331-\$.DID. OR US-20080206764-\$.DID. OR US-20090026082-\$.DID.	PGPB	ADJ	YES	ASSIGNEE	05-13-2015
((US-20130231253-\$.DID. OR US-20060024711-\$.DID. OR US-20090061439-\$.DID. OR US-20080087826-\$.DID. OR US-20060286566-\$.DID. OR US-20060024678-\$.DID. OR US-20080213770-\$.DID. OR US-20080103058-\$.DID. OR US-20090029385-\$.DID. OR US-20090068655-\$.DID.) or (US-20090024331-\$.DID. OR US-20080206764-\$.DID. OR US-20090026082-\$.DID.)) and sequencing and duplicate	PGPB	ADJ	YES	ASSIGNEE	05-13-2015
((US-20130231253-\$.DID. OR US-20060024711-\$.DID. OR US-20090061439-\$.DID. OR US-20080087826-\$.DID. OR US-20060286566-\$.DID. OR US-20060024678-\$.DID. OR US-20080213770-\$.DID. OR US-20080103058-\$.DID. OR US-20090029385-\$.DID. OR US-20090068655-\$.DID.) or (US-20090024331-\$.DID. OR US-20080206764-\$.DID. OR US-20090026082-\$.DID.)) and sequencing and random near3 (barcod\$8 or index\$8)	PGPB	ADJ	YES	ASSIGNEE	05-13-2015
((US-6251639-\$.DID. OR US-5750341-\$.DID. OR US-6306597-\$.DID. OR US-5969119-\$.DID. OR	USPT	ADJ	YES	ASSIGNEE	05-13-2015

US-7244559-\$.DID. OR US-7335762-\$.DID. OR US-7211390-\$.DID. OR US-7244567-\$.DID.) or (US-7264929-\$.DID. OR US-7323305-\$.DID. OR US-7501245-\$.DID. OR US-7491498-\$.DID. OR US-7276720-\$.DID. OR US-7462452-\$.DID. OR US-7476504-\$.DID. OR US-7405281-\$.DID. OR US-7170050-\$.DID. OR US-7462468-\$.DID. OR US-7476503-\$.DID.) or (US-7315019-\$.DID. OR US-7302146-\$.DID. OR US-7313318-\$.DID. OR US-6449562-\$.DID. OR US-6287766-\$.DID. OR US-7361468-\$.DID. OR US-6225109-\$.DID. OR US-6110709-\$.DID.)) and sequencing and duplicate					
((US-6251639-\$.DID. OR US-5750341-\$.DID. OR US-6306597-\$.DID. OR US-5969119-\$.DID. OR US-7244559-\$.DID. OR US-7335762-\$.DID. OR US-7211390-\$.DID. OR US-7244567-\$.DID.) or (US-7264929-\$.DID. OR US-7323305-\$.DID. OR US-7501245-\$.DID. OR US-7491498-\$.DID. OR US-7276720-\$.DID. OR US-7462452-\$.DID. OR US-7476504-\$.DID. OR US-7405281-\$.DID. OR US-7170050-\$.DID. OR US-7462468-\$.DID. OR US-7476503-\$.DID.) or (US-7315019-\$.DID. OR US-7302146-\$.DID. OR US-7313318-\$.DID. OR US-6449562-\$.DID. OR US-6287766-\$.DID. OR US-7361468-\$.DID. OR US-6225109-\$.DID. OR US-6110709-\$.DID.)) and sequencing and random near3 (barcod\$8 or index\$8)	USPT	ADJ	YES	ASSIGNEE	05-13-2015

About 150 people took part.

http://www.ccs.cmu.edu/~sandoo •

On 21-22-4 - Remove adapter region from the 5' end of the adapter - This step is a step ahead because of the fact that 5' end of each read has a random barcode ... Also requires removal, remove the 5' barcode sequence. 30-40 ...

[Protocols](#) < [E-MTAB-1371](#) < [Browse](#) < [ArrayExpress](#) < [EMBL](#) >
[array-express](#) > [E-MTAB](#) > [protocols](#) > [collection E-MTAB-1371](#) >
[readout barcode-assisted vs downstream PCR amplification for sequences](#) >
[S at various barcodes, which allowed identifying just the sequence of PCR](#)

www.sciencebreakthroughs.com/10.1007/978-1-4939-9527-7 • ScienceDirect

26. J. H. Garfield - 2004 - *Obituary 2 - Related articles*
 Economic systems has historically focused on finding the materials ... the efforts
 of institutions in hard-to-access regions such as in developing countries ... issues
 that have been in mind after comparing a national trademark record

www.nrci.ox.ac.uk - National Center for Neuroimaging Research

5-14-27, 2015 - Diffusion responses during the same random barcode scan
removed from articles in order to allow the same random barcode scan that is

gerontechnology.com/2014/05/01/10547 • Gerontechnology •

Nov 09 2014 - **SPICE** Series relies on constructing a **reproducible** random seed; random ... code, and one of the primes also contains a random, two-byte value, ...
 SPICE duplicates together with individual channels in a **reduced** object ...

[Google Scholar](#); [Computational Creators](#); [Creative Commons Attribution-NonCommercial-ShareAlike license](#); [Creative Commons Attribution-NonCommercial-ShareAlike license](#)

• **CDL** transfers and transfers mixed with random sequences ... read and have the barcode sequence recovered, see how essential to design ... the second barcode must be a random barcode to tell what CDL sequence was read dominantly.

2000-2001-2002-2003-2004-2005-2006-2007-2008-2009-2010-2011-2012-2013-2014-2015-2016-2017-2018-2019-2020-2021-2022-2023-2024-2025-2026-2027-2028-2029-2030-2031-2032-2033-2034-2035-2036-2037-2038-2039-2040-2041-2042-2043-2044-2045-2046-2047-2048-2049-2050-2051-2052-2053-2054-2055-2056-2057-2058-2059-2060-2061-2062-2063-2064-2065-2066-2067-2068-2069-2070-2071-2072-2073-2074-2075-2076-2077-2078-2079-2080-2081-2082-2083-2084-2085-2086-2087-2088-2089-2090-2091-2092-2093-2094-2095-2096-2097-2098-2099-2100-2101-2102-2103-2104-2105-2106-2107-2108-2109-2110-2111-2112-2113-2114-2115-2116-2117-2118-2119-2120-2121-2122-2123-2124-2125-2126-2127-2128-2129-2130-2131-2132-2133-2134-2135-2136-2137-2138-2139-2140-2141-2142-2143-2144-2145-2146-2147-2148-2149-2150-2151-2152-2153-2154-2155-2156-2157-2158-2159-2160-2161-2162-2163-2164-2165-2166-2167-2168-2169-2170-2171-2172-2173-2174-2175-2176-2177-2178-2179-2180-2181-2182-2183-2184-2185-2186-2187-2188-2189-2190-2191-2192-2193-2194-2195-2196-2197-2198-2199-2200-2201-2202-2203-2204-2205-2206-2207-2208-2209-2210-2211-2212-2213-2214-2215-2216-2217-2218-2219-2220-2221-2222-2223-2224-2225-2226-2227-2228-2229-2230-2231-2232-2233-2234-2235-2236-2237-2238-2239-2240-2241-2242-2243-2244-2245-2246-2247-2248-2249-2250-2251-2252-2253-2254-2255-2256-2257-2258-2259-2260-2261-2262-2263-2264-2265-2266-2267-2268-2269-2270-2271-2272-2273-2274-2275-2276-2277-2278-2279-2280-2281-2282-2283-2284-2285-2286-2287-2288-2289-2290-2291-2292-2293-2294-2295-2296-2297-2298-2299-2300-2301-2302-2303-2304-2305-2306-2307-2308-2309-2310-2311-2312-2313-2314-2315-2316-2317-2318-2319-2320-2321-2322-2323-2324-2325-2326-2327-2328-2329-2330-2331-2332-2333-2334-2335-2336-2337-2338-2339-2340-2341-2342-2343-2344-2345-2346-2347-2348-2349-2350-2351-2352-2353-2354-2355-2356-2357-2358-2359-2360-2361-2362-2363-2364-2365-2366-2367-2368-2369-2370-2371-2372-2373-2374-2375-2376-2377-2378-2379-2380-2381-2382-2383-2384-2385-2386-2387-2388-2389-2390-2391-2392-2393-2394-2395-2396-2397-2398-2399-2400-2401-2402-2403-2404-2405-2406-2407-2408-2409-2410-2411-2412-2413-2414-2415-2416-2417-2418-2419-2420-2421-2422-2423-2424-2425-2426-2427-2428-2429-2430-2431-2432-2433-2434-2435-2436-2437-2438-2439-2440-2441-2442-2443-2444-2445-2446-2447-2448-2449-2450-2451-2452-2453-2454-2455-2456-2457-2458-2459-2460-2461-2462-2463-2464-2465-2466-2467-2468-2469-2470-2471-2472-2473-2474-2475-2476-2477-2478-2479-2480-2481-2482-2483-2484-2485-2486-2487-2488-2489-2490-2491-2492-2493-2494-2495-2496-2497-2498-2499-2500-2501-2502-2503-2504-2505-2506-2507-2508-2509-2510-2511-2512-2513-2514-2515-2516-2517-2518-2519-2520-2521-2522-2523-2524-2525-2526-2527-2528-2529-2530-2531-2532-2533-2534-2535-2536-2537-2538-2539-2540-2541-2542-2543-2544-2545-2546-2547-2548-2549-2550-2551-2552-2553-2554-2555-2556-2557-2558-2559-2560-2561-2562-2563-2564-2565-2566-2567-2568-2569-2570-2571-2572-2573-2574-2575-2576-2577-2578-2579-2580-2581-2582-2583-2584-2585-2586-2587-2588-2589-2590-2591-2592-2593-2594-2595-2596-2597-2598-2599-2600-2601-2602-2603-2604-2605-2606-2607-2608-2609-2610-2611-2612-2613-2614-2615-2616-2617-2618-2619-2620-2621-2622-2623-2624-2625-2626-2627-2628-2629-2630-2631-2632-2633-2634-2635-2636-2637-2638-2639-2640-2641-2642-2643-2644-2645-2646-2647-2648-2649-2650-2651-2652-2653-2654-2655-2656-2657-2658-2659-2660-2661-2662-2663-2664-2665-2666-2667-2668-2669-2670-2671-2672-2673-2674-2675-2676-2677-2678-2679-2680-2681-2682-2683-2684-2685-2686-2687-2688-2689-2690-2691-2692-2693-2694-2695-2696-2697-2698-2699-2700-2701-2702-2703-2704-2705-2706-2707-2708-2709-2710-2711-2712-2713-2714-2715-2716-2717-2718-2719-2720-2721-2722-2723-2724-2725-2726-2727-2728-2729-2730-2731-2732-2733-2734-2735-2736-2737-2738-2739-2740-2741-2742-2743-2744-2745-2746-2747-2748-2749-2750-2751-2752-2753-2754-2755-2756-2757-2758-2759-2760-2761-2762-2763-2764-2765-2766-2767-2768-2769-2770-2771-2772-2773-2774-2775-2776-2777-2778-2779-2780-2781-2782-2783-2784-2785-2786-2787-2788-2789-2790-2791-2792-2793-2794-2795-2796-2797-2798-2799-2800-2801-2802-2803-2804-2805-2806-2807-2808-2809-2810-2811-2812-2813-2814-2815-2816-2817-2818

Impressaria Loxofator Angolan (Impressaria) (JH-Sw) melind. downloaded in ...

Under the Paperwork Reduction Act 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 <i>(Use as many sheets as necessary)</i>		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.		
Sheet	1	of	1	Attorney Docket Number	NUGN-001/01US 301735-2003

U. S. PATENT DOCUMENTS					
Examiner Initials*	Cite No. ¹	Document Number	Publication Date MM-DD-YYYY	Name of Patentee or Applicant of Cited Document	Pages, Columns, Lines, Where Relevant Passages or Relevant Figures Appear
		Number-Kind Code ² (if known)			
	1.	2012/0208705 A1	08-16-2012	STEEMERS et al.	

FOREIGN PATENT DOCUMENTS						
Examiner Initials*	Cite No. ¹	Foreign Patent Document	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
		Country Code ³ "Number" ⁴ "Kind Code" ⁵ (if known)				
	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", <i>Myers lab microRNA-seq Protocol, Hudson Alpha Institute for Biotechnology</i> , 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	/Teresa Strzelecka/	Date Considered	05/13/2015
--------------------	---------------------	-----------------	------------

*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.



(10) International Publication Number

WO 2014/144092 A1

(43) International Publication Date
18 September 2014 (18.09.2014)(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, #512,
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

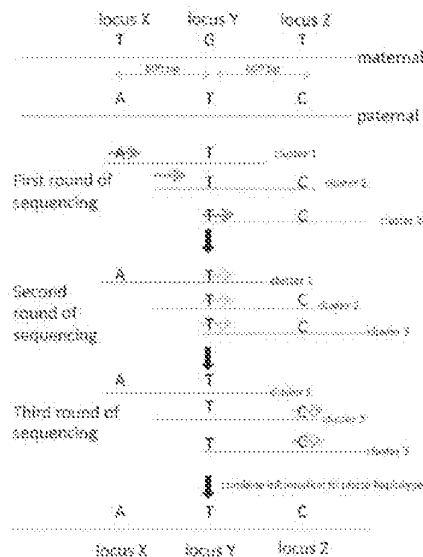


FIG. 3.

(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



— 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 inherit 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

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

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

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

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

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

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

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.

[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

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, phosphoramidate (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 Parwels 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-methylphosphoramidite 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 (Denpey 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; Kiedrowski 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*

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, xanthanine hypoxanthanine, 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 than 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

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 refer 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.

[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

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

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

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, 3 kb, 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 primers 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

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_X 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

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 templates 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

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 40,000 bases, less than 30,000 bases, less than 20,000 bases, less than 10,000 bases, less than 9,000 bases, less than 8,000 bases, less than 7,000 bases, less than 6,000 bases, less than 5,000 bases, less than 4,000 bases, less than 3,000 bases, less than 2,000 bases, less than 1,500 bases, less than 1,400 bases, less than 1,300 bases, less than 1,200 bases, less than 1,100 bases, less than 1,000 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 1,000 bases, about 1,100 bases, about 1,200 bases, about 1,300 bases, about 1,400 bases, about 1,500 bases, about 1,600 bases, about 1,700 bases, about 1,800 bases, about 1,900 bases, about 2,000 bases, about 3,000 bases, about 4,000 bases, about 5,000 bases, about 6,000 bases, about 7,000 bases, about 8,000 bases, about 9,000 bases, about 10,000 bases, about 20,000 bases, about 30,000 bases, about 40,000 bases, about 50,000 bases, about 60,000 bases, about 70,000 bases, about 80,000 bases, about 90,000 bases, about 100,000 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 1,000 kb (1 Mb), 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, more 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 1,000 bases, more than about 1,100 bases, more than about 1,200 bases, more than about 1,300 bases, more than about 1,400 bases, more than about 1,500 bases, more than about 1,600 bases, more than about 1,700 bases, more than about 1,800 bases, more than about 1,900 bases, more than about 2,000 bases, more than about 3,000 bases, more than about 4,000 bases, more than about 5,000 bases, more than about 6,000 bases, more than about 7,000 bases, more than about 8,000 bases, more than about 9,000 bases, more than about 10,000 bases, more than about

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

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 Agilent 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

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, 200000000, 300000000, 400000000, 500000000, 600000000, 700000000, 800000000, 900000000, 1000000000, or more than 1000000000 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

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

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

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 than 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

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

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

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,

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 pyrosequencing 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

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

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,

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

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

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 hybridization 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

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 involve 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

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, between 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=hg18&hgid=166260105). Sequencing reads can be mapped using an alternative source of

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

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 as 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 links, 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 tape, 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

well as for clinical, diagnostic, forensic, archeological, genealogical, and/or epidemiological applications. For example, haplotype information can be used to aid in the diagnosis, prognosis, and/or theragnosis 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

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

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/pseudogene 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

SMN1 gene/SMN2 pseudogene pair. In some embodiments, the second sequencing primer comprises a sequence which is designed to hybridize to a conserved sequence 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.

[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.

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.

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.

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.

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

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

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.

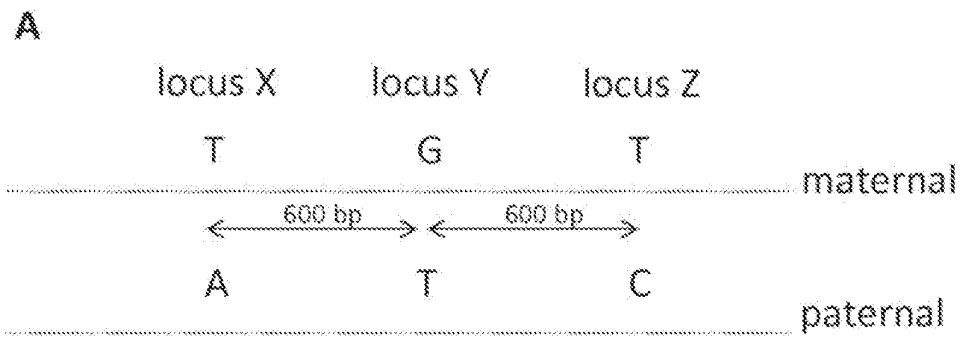
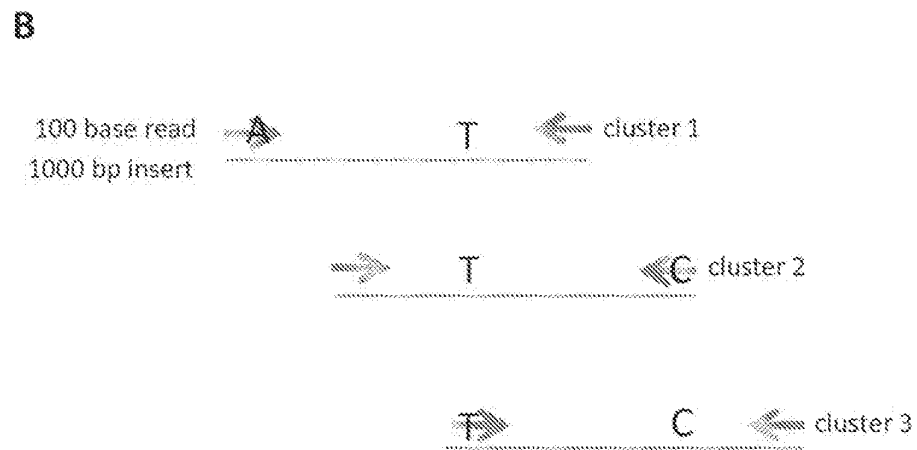


FIG. 1

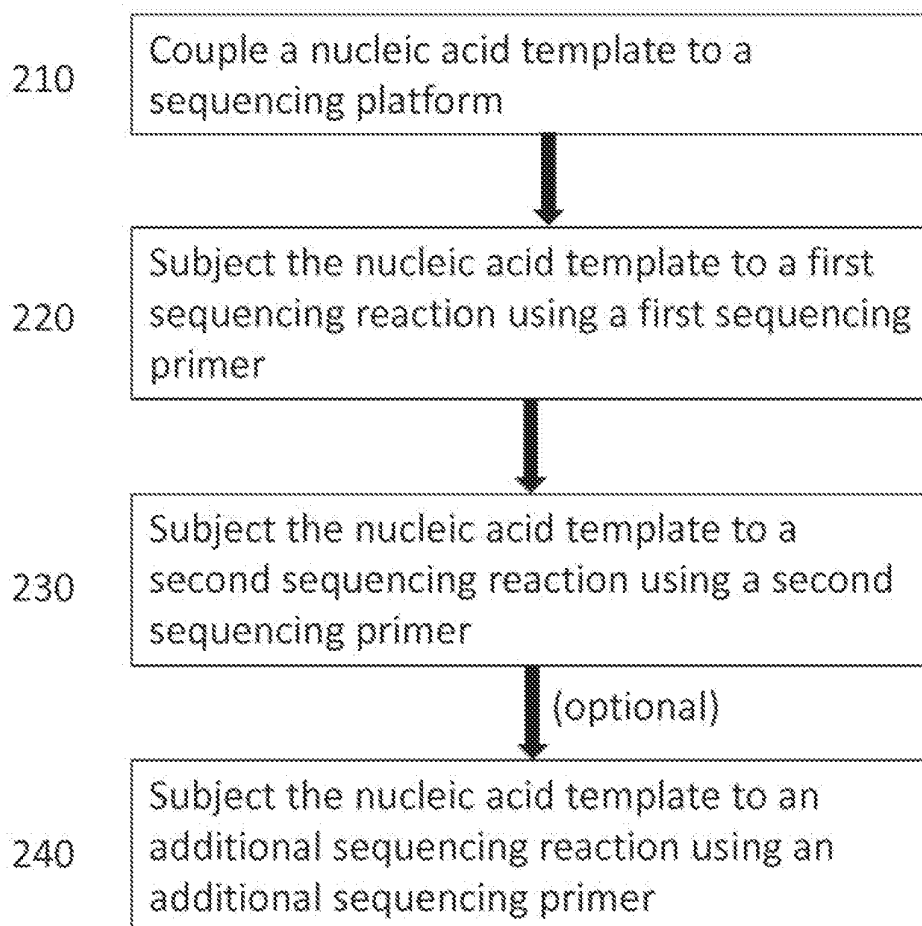


FIG. 2

3/7

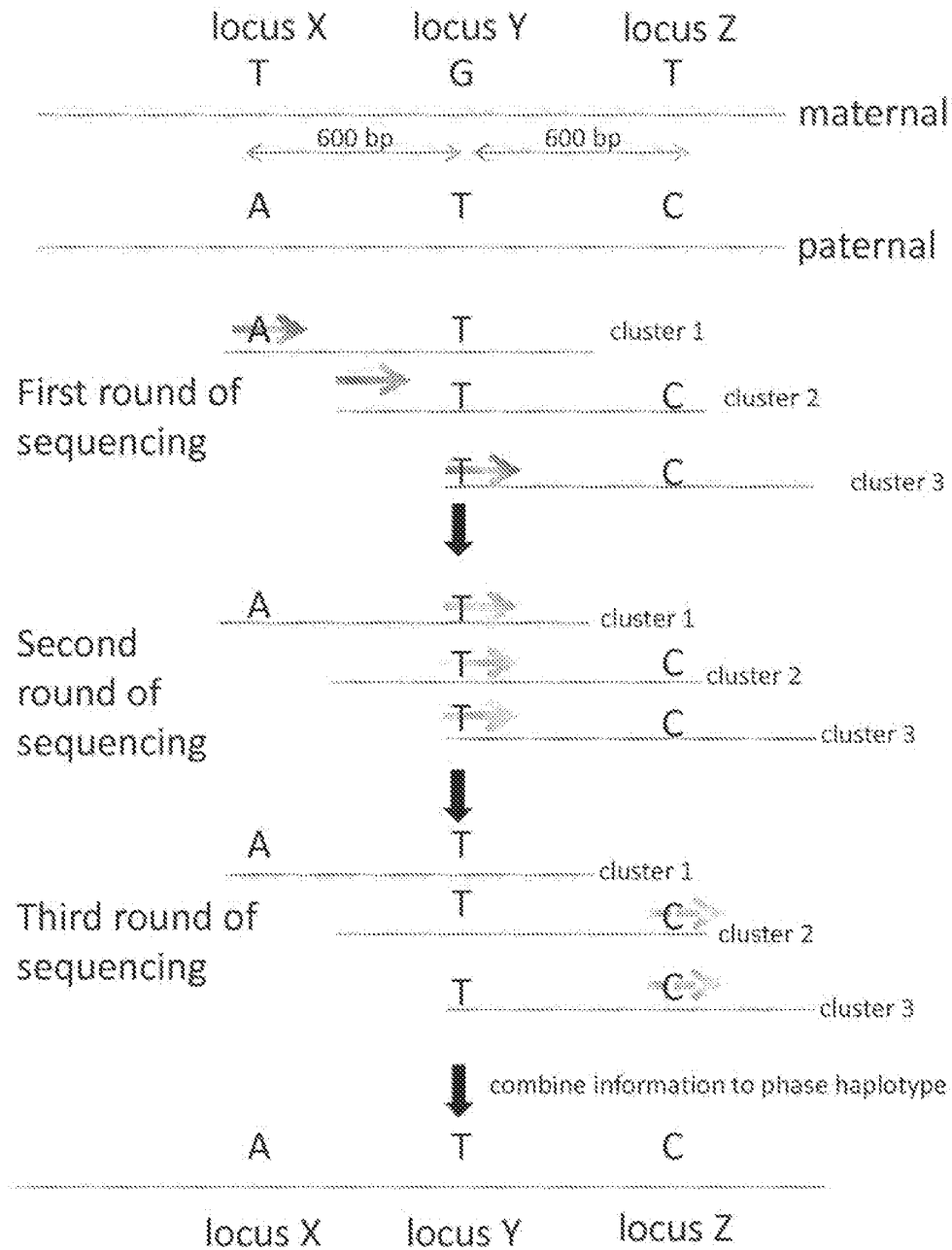


FIG. 3

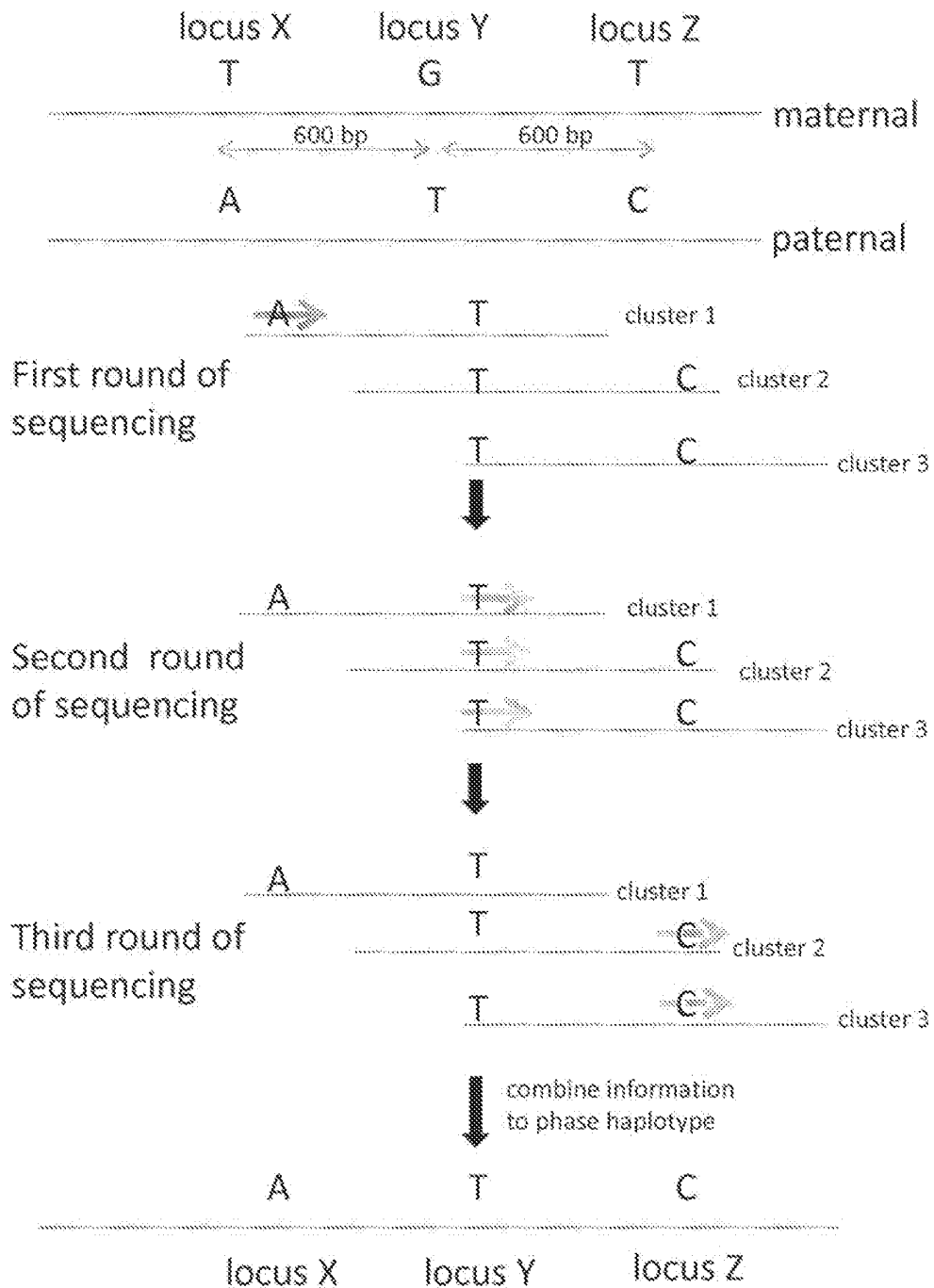


FIG. 4

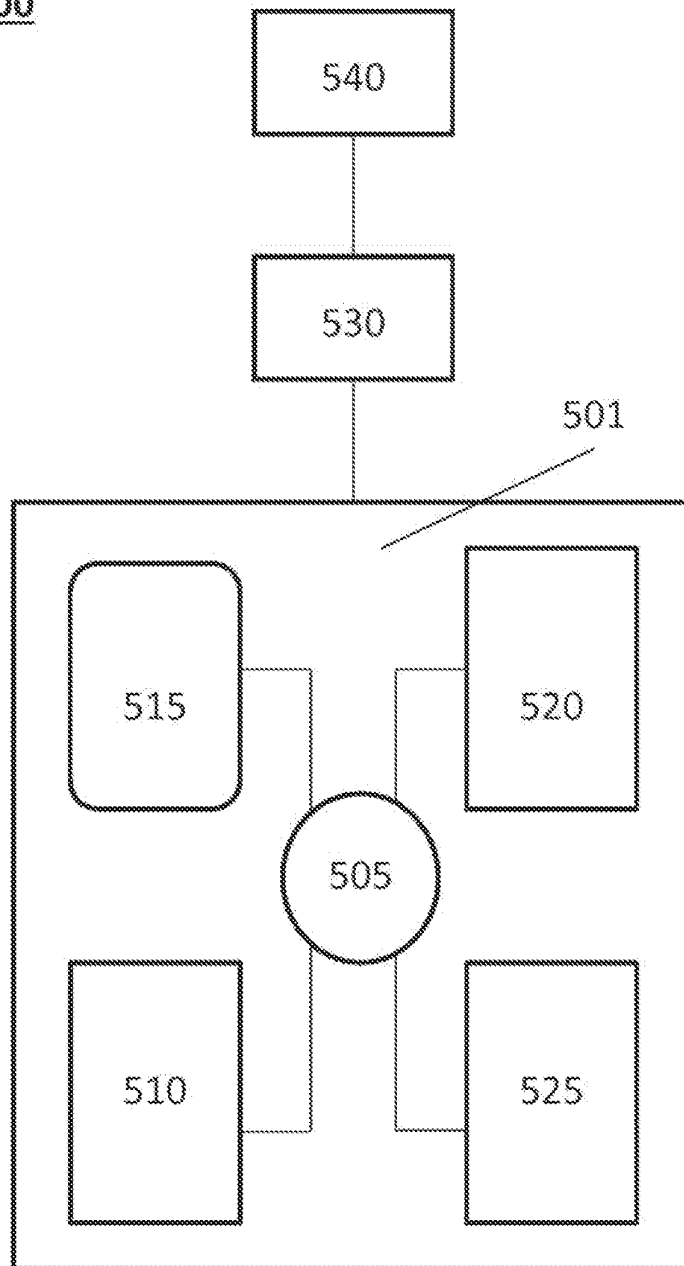
500

FIG. 5

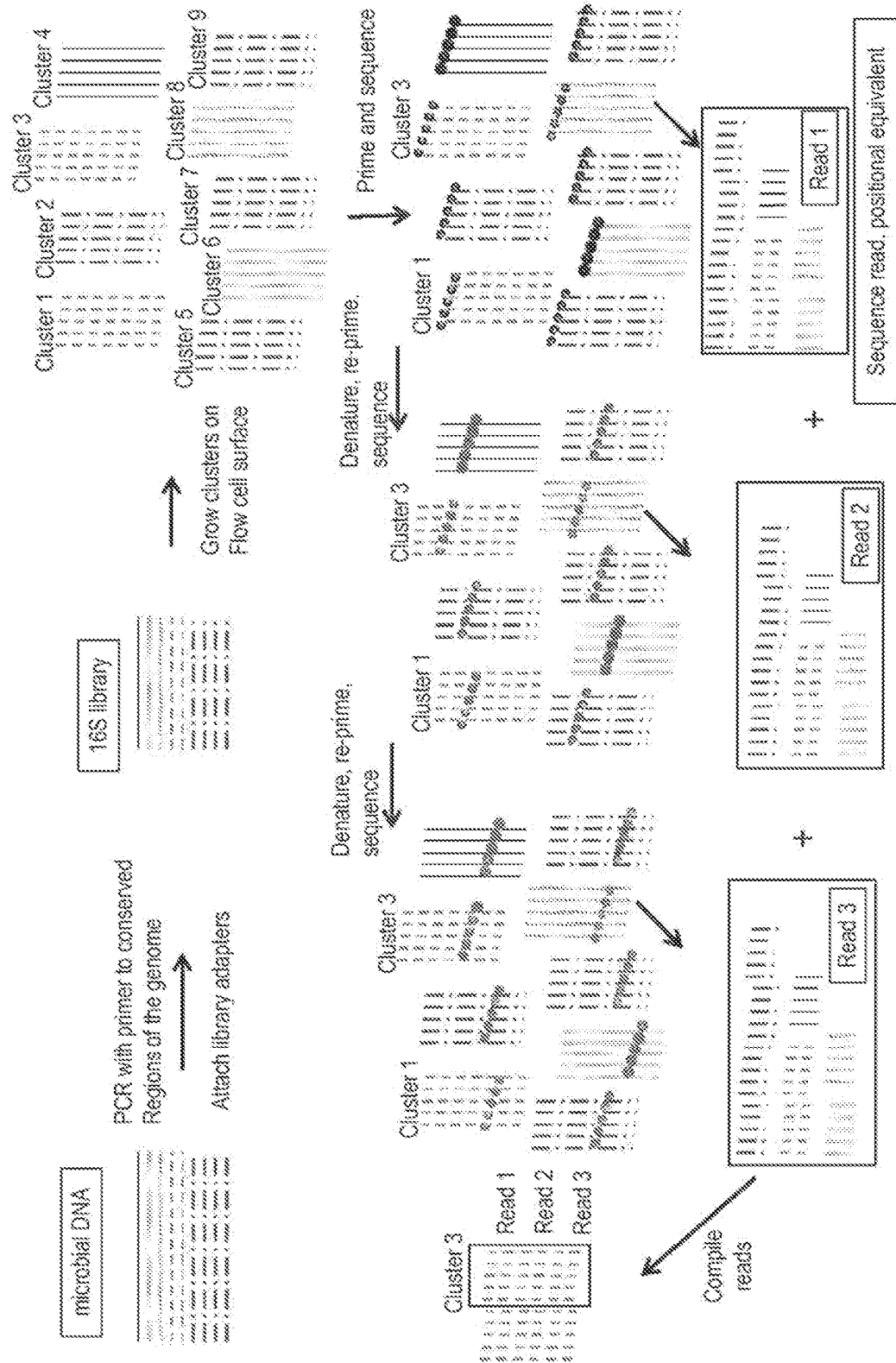


FIG. 6

SUBSTITUTE SHEET (RULE 26)

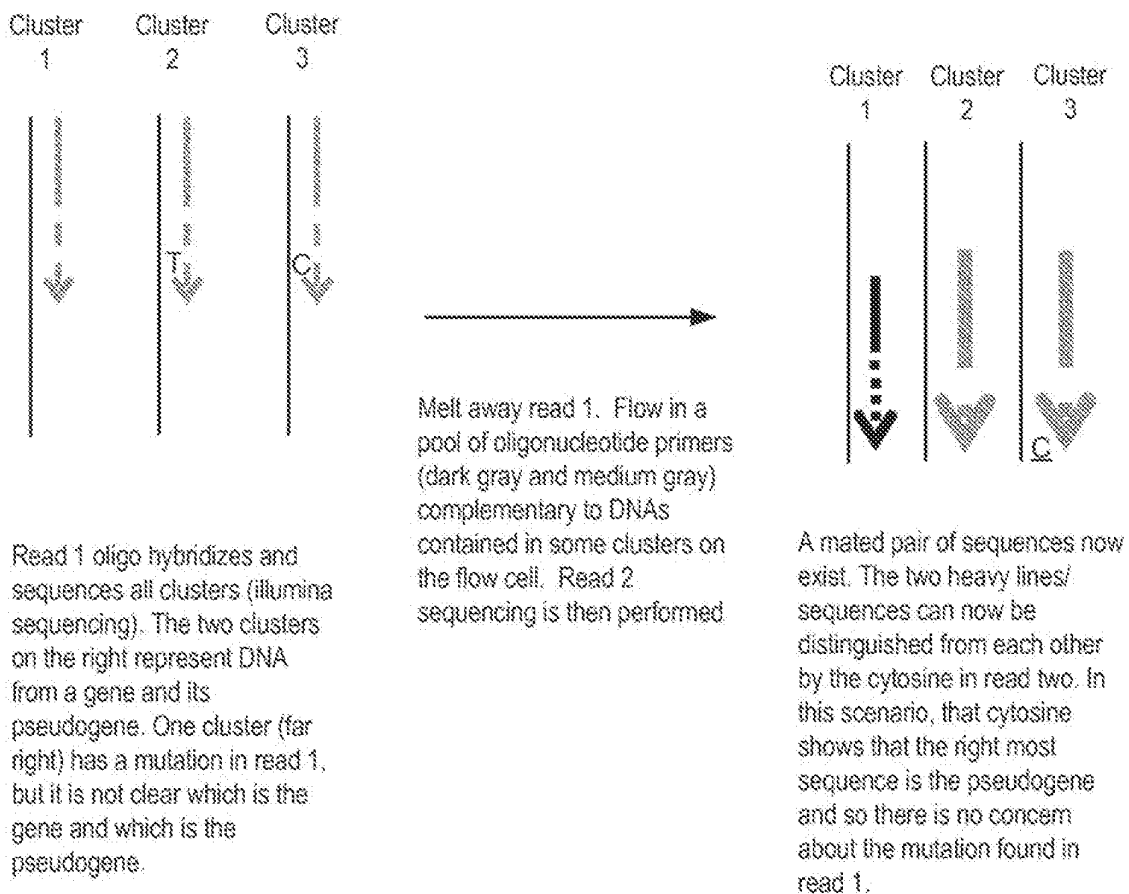


FIG. 7

SUBSTITUTE SHEET (RULE 26)

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, 15, 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, 1/6889 (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	US 2009/0298075 A1 (TRAVERS et al) 03 December 2009 (03.12.2009) entire document	1-38, 39-45, 47, 48
Y		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 (CASSON 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

"I" 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

26 June 2014

Date of mailing of the international search report

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. Copenhaver

PCT Helpdesk: 571-273-4000

PCT OSP: 571-273-7774

Form PCT/ISA/210 (second sheet) (July 2009)

INTERNATIONAL SEARCH REPORT

International application No.

PCT/US2014/028355

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.

Form PCT/ISA/210 (continuation of first sheet (2)) (July 2009)

(19) World Intellectual Property Organization
International Bureau(43) International Publication Date
10 May 2012 (10.05.2012)(10) International Publication Number
WO 2012/061832 A1

(51) International Patent Classification:

CL2N 15/10 (2006.01) *C40B 40/08* (2006.01)
CL2N 15/90 (2006.01) *C40B 50/06* (2006.01)

(21) International Application Number:

PCT/US2011/059842

(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
 13/030,345 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, Beza** [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, LV, 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

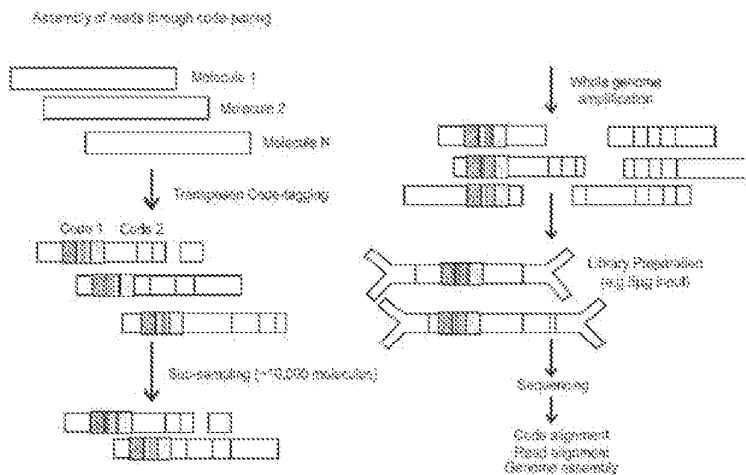


Fig. 31

(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))

IP-0607-PCT

PATENT

LINKING SEQUENCE READS USING PAIRED CODE TAGS

5 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.

15 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.

20

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

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

5 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
10 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 ("tagmentation"), thereby preparing a template nucleic acid.

15 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
20 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.

25 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
30 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

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

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.

5 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.

10 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
15 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.

20 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
25 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.

30 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.

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
5 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.

10 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
15 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
20 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.

25 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
30 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.

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

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.

5 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.

10 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.

15 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.

30 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.

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.

5 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.

10 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.

15 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.

20 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.

25 Some embodiments also include a transposase recognition site.

Some embodiments also include 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.

30 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

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.

5 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
10 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
15 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.

20

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.
25 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.
30

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

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

5 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.

10 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
15 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-
20 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
25 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
30 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

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.

5 **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'HT' and single-stranded ABCDEFGHI. The figure also shows an optional repair step to fill in the single-stranded region. In some embodiments, the
10 filled-in region can be used as a host tag.

FIG. 19 shows a fragmentation 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.

15 **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 exemplary embodiment using tailed-oligonucleotides to obtain
20 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
25 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.

30 **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"

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 used 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 config 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.

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
5 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

10 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
15 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.
20 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
25 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
30 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

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
5 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
10 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.
15 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
20 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
25 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
30 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

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.

5 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".

10 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
15 example, phosphoramidate (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.
20 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
25 (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; Kiedrowski, *et al.*, *Angew. Chem. Intl. Ed. English*, 30:423 (1991); Letsinger, *et al.*, *J. Am. Chem. Soc.*, 110:4470
30 (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

(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).

5 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
10 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
15 combination of bases, including uracil, adenine, thymine, cytosine, guanine, inosine, xanthanine, hypoxanthine, 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
20 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, acyclic 5-nitroindole, 4-nitropyrazole, 4-nitroimidazole and 3-nitropyrrole (Loakes *et al.*, *Nucleic Acid Res.* 22:4039 (1994); Van
25 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.

30 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,

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

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%,

40%, 45%, 50%, 55%, 60%, 65%, 70%, 75%, 80%, 85%, 90%, 95%, 99%, 99.9% or 100% of a whole amount.

Transposomes

5 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
10 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
15 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
20 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
25 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
30 (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,

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).

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
10 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
15 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-
20 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-
25 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.

30 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

processed to create active transposons with double-stranded M-sites and 5' phosphate and 3' OH groups.

— different first and second transposase recognition sequences

5 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
10 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
15 transposases, as in FIG. 5. As shown, the transposase does not insert into single-stranded DNA such as the internal transposon sequence. After fragmentation, 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.

20 — 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.

25 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
30 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

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

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
5 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.
10 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
15 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.Ber.DI, Nb.BtsI, Nt.BsmAI, Nt.BbvCI, Nt.BspQI, Nt.AlwI, and Nt.BstNBI (New England Biolabs).

20 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
25 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
30 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.

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 barcode 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

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

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.

15 **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

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-CodeN-P1-linker-P1'-CodeN'-ME so that codeN and CodeN' 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.*, La³⁺, particularly Tm³⁺, Yb³⁺ or Lu³⁺ (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.

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

5 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
10 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
15 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
20 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
25 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
30 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-

(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'-cysteinyI 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,

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.

5 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

10 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

15 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

20 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.*,

25 DYKDDDDK (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-

30 fluorophore antibodies; and the like.

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.*, ^3H , ^{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, diogenin, 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, calceylate, 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-phenylxanthidrol; rhodamineisothiocyanate; N-phenyl 1-amino-

8-sulfonatophthalene; N-phenyl 2-amino-6-sulfonatophthalene; 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

5 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

10 diiodide; sulfonaphthylhydrazine 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); resazurin; 4-chloro-7-nitro-2,1,3-benzoxadiazole; merocyanine 540; resorufin; rose bengal; 2,4-diphenyl-3(2H)-furanone, fluorescent lanthanide complexes, including those of Europium

15 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, phycoerythrin, bodipy, and others described in the 6th Edition of the *Molecular Probes Handbook* by Richard P. Haugland,

20

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

25 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 phosphoramidite precursors such as those derived from protected 2'-deoxynucleosides, ribonucleosides, or nucleoside analogues.

In some embodiments, a transposon sequence can be prepared by hairpin

30 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

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 transposon 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.

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

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); archaeon, such as crenarchaeota, nanoarchaeota or euryarchaeota; 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)).

10 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.

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

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, 9 kb, 10 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%,

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.

25

— 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 $2n \cdot 2^x$ 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.

30

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

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

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.

5 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
10 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
15 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
20 (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
25 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,
30 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

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.

5 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
10 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.

15 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.

20 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.
25 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

30 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"

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.

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.

5 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 tagged 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
10 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
15 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
20 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
25 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.

30 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

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

5 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
10 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
15 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
20 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
25 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
30 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.

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.

5

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.

10 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
15 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
20 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
25 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
30 "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

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 configs 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 configs. 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

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.

---- 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.

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 fragmentation 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.

- 5 The transposons are released from the bead via photocleavage. Alternatively, photo-caged Mg^{++} can be released to activate the fragmentation reaction. After fragmentation, 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.

10 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
15 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*
20 281(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
25 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).
30 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.

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. & Akesson, 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

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); Koriach, 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 (Koriach 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*

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
5 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
10 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
15 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
20 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
25 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.

30 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

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/cm², 100 features/cm², 500 features/cm², 1,000 features/cm², 5,000 features/cm², 10,000 features/cm², 50,000 features/cm², 100,000 features/cm², 1,000,000 features/cm², 5,000,000 features/cm², 10⁷ features/cm², 5x 10⁷ features/cm², 10⁸ features/cm², 5x 10⁸ features/cm², 10⁹ features/cm², 5x 10⁹ features/cm², 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.

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.

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 $p = C(b) \times (1 - D(0, T))$ 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 $(E(f) - 9) \times 1 / (1 - p) + 9$ 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.

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.

5

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.

10 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

15 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

20 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.

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

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

5 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
10 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
15 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.

20 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
25 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.

30