(19) DANMARK

(10) **DK/EP 2970951 T3**

Patent- og Varemærkestyrelsen

(12) Oversættelse af europæisk patentskrift

(51) Int.Cl.: *C 12 Q 1/6869 (2018.01)* *C 12 N 15/10 (2006.01)*

(45) Oversættelsen bekendtgjort den: **2019-05-13**

(80) Dato for Den Europæiske Patentmyndigheds bekendtgørelse om meddelelse af patentet: **2019-02-20**

(86) Europæisk ansøgning nr.: **13711276.9**

(86) Europæisk indleveringsdag: **2013-03-13**

(87) Den europæiske ansøgnings publiceringsdag: **2016-01-20**

(86) International ansøgning nr.: **US2013031023**

(87) Internationalt publikationsnr.: **WO2014142850**

(84) Designerede stater: **AL AT BE BG CH CY CZ DE DK EE ES FI FR GB GR HR HU IE IS IT LI LT LU LV MC MK MT NL NO PL PT RO RS SE SI SK SM TR**

(73) Patenthaver: **Illumina, Inc., 5200 Ilumina Way, San Diego, CA 92122, USA**

(72) Opfinder: **STEEMERS, Frank, 5200 Illumina Way, San Diego, CA 92122, USA**
AMINI, Sasan, 5200 Illumina Way, San Diego, CA 92122, USA
GUNDERSON, Kevin, 1543 Juniper Hill Drive, Encinitas, CA 92024, USA
GORYSHIN, Igor, 1013 Greenwich Drive, Madison, WI 53711, USA
PIGNATELLI, Natasha, 15195 El Camino Grande, Saratoga, CA 95070, USA

(74) Fuldmægtig i Danmark: **HEGNER & PARTNERS A/S, Symbion Science Park, Fruebjergvej 3, 2100 København Ø, Danmark**

(54) Benævnelse: **FREMGANGSMÅDER TIL NUKLEINSYRESEKVENTERING**

(56) Fremdragne publikationer:
WO-A1-2012/061832
WO-A1-2012/103545
US-A1- 2012 208 705
US-A1- 2012 208 724
HAOYI WANG ET AL: "Calling Cards enable multiplexed identification of the genomic targets of DNA-binding proteins", GENOME RESEARCH, COLD SPRING HARBOR LABORATORY PRESS, UNITED STATES, vol. 21, no. 5, 1 May 2011 (2011-05-01), pages 748-755, XP002672204, ISSN: 1549-5469, DOI: 10.1101/GR.114850.110 [retrieved on 2011-04-06]
SYED FRAZ ET AL: "Optimized library preparation method for next-generation sequencing", NATURE METHODS, NATURE PUBLISHING GROUP, GB, vol. 6, no. 10, 1 October 2009 (2009-10-01), pages I-II, XP002672202, ISSN: 1548-7091
RACHEL MARINE ET AL: "Evaluation of a transposase protocol for rapid generation of shotgun high-throughput sequencing libraries from nanogram quantities of DNA", APPLIED AND ENVIRONMENTAL MICROBIOLOGY, AMERICAN SOCIETY FOR MICROBIOLOGY, US, vol. 77, no. 22, 1 November 2011 (2011-11-01), pages 8071-8079, XP002672205, ISSN: 1098-5336, DOI: 10.1128/AEM.05610-11 [retrieved on 2011-09-23]
CARUCCIO NICHOLAS ED - CHENG QIONG: "Preparation of next-generation sequencing libraries using

Fortsættes ...

Nextera(TM) technology: simultaneous DNA fragmentation and adaptor tagging by in vitro transposition", METHODS IN MOLECULAR BIOLOGY; [METHODS IN MOLECULAR BIOLOGY ; 834], HUMANA PRESS, US, vol. 733, 1 January 2011 (2011-01-01), pages 241-255, XP009157814, ISSN: 1940-6029, DOI: 10.1007/978-1-61779-089-8_17 [retrieved on 2011-01-01]

DESCRIPTION

FIELD OF THE INVENTION

[0001] Embodiments of the present disclosure relate to sequencing nucleic acids. In particular, embodiments of the methods and compositions provided herein relate to preparing nucleic acid templates and obtaining sequence data therefrom.

BACKGROUND OF THE INVENTION

[0002] 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.

[0003] 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

[0004] Some embodiments of the methods and compositions provided herein include a method of obtaining sequence information from a target nucleic acid, said method comprising: (a) obtaining a template nucleic acid comprising a plurality of transposomes inserted into said target nucleic acid, wherein at least some of the inserted transposomess each comprise a first transposon sequence, a second transposon sequence noncontiguous with said first transposon sequence, and a transposase associated with the first transposon sequence and the second transposon sequence; (b) compartmentalizing the template nucleic acid comprising said plurality of inserted transposomes into each vessel of a plurality of vessels; (c) removing the transposase from the template nucleic acid; and (d) obtaining sequence information from the template nucleic acid of each vessel.

[0005] In some embodiments, step (b) comprises providing each vessel with an amount of template nucleic acid equal to about less than one haploid equivalent, about equal than one haploid equivalent, or more than one haploid equivalent of the target nucleic acid.

[0006] In some embodiments, step (b) comprises providing each vessel with an amount of template nucleic acid less than about one haploid equivalent of the target nucleic acid.

[0007] In some embodiments, step (c) comprises a method selected from the group consisting of adding a detergent, changing temperature, changing pH, adding a protease, adding a chaperone, and adding a polymerase.

[0008] In some embodiments, the first transposon sequence comprises a first primer site and the second transposon sequences comprise a second primer site.

[0009] In some embodiments, the first primer site further comprises a first barcode and the second primer site further comprises a second barcode.

[0010] In some embodiments, the first barcode and second barcode are different.

[0011] In some embodiments, the target nucleic acid comprises an amplified nucleic acid.

[0012] In some embodiments, the target nucleic acid is obtained by enriching a plurality of nucleic acids for a sequence of interest before or after transposition.

[0013] In some embodiments, step (a) further comprises enriching the template nucleic acid for a sequence of interest.

[0014] In some embodiments, the target nucleic acid comprises genomic DNA.

[0015] In some embodiments, step (d) further comprises assembling from sequence data a representation of at least a portion of said template nucleic acid from each vessel.

[0016] In some embodiments, the sequence information comprises haplotype sequence information.

[0017] Some embodiments of the methods and compositions provided herein include a method for preparing a library of template nucleic acids to obtain sequence information from a target nucleic acid, said method comprising: (a) preparing a template nucleic acid comprising a plurality of transposomes inserted into said target nucleic acid, wherein at least some of the inserted transposome each comprise a first transposon sequence, a second transposon sequence noncontiguous with said first transposon sequence, and a transposase associated with the first transposon sequence and the second transposon sequence; and (b) compartmentalizing the template nucleic acid comprising said plurality of inserted transposomes into each vessel of a plurality of vessels; and (c) removing the transposase from the template nucleic acid.

[0018] In some embodiments, step (b) comprises providing each vessel with an amount of template nucleic acid equal to less than one haploid equivalent, about one equivalent, or more

than one equivalent of the target nucleic acid.

[0019] In some embodiments, step (b) comprises providing each vessel with an amount of template nucleic acid less than about one haploid equivalent of the target nucleic acid.

[0020] In some embodiments, step (c) comprises a method selected from the group consisting of adding a detergent, changing temperature, changing pH, adding a protease, adding a chaperone, and adding a polymerase.

[0021] In some embodiments, the first transposon sequence comprises a first primer site and the second transposon sequences comprise a second primer site.

[0022] In some embodiments, the first primer site further comprises a first barcode and the second primer site further comprises a second barcode.

[0023] In some embodiments, the first barcode and second barcode are different.

[0024] In some embodiments, the target nucleic acid comprises an amplified nucleic acid.

[0025] In some embodiments, the target nucleic acid is obtained by enriching a plurality of nucleic acids for a sequence of interest.

[0026] In some embodiments, step (a) further comprises enriching the template nucleic acid for a sequence of interest.

[0027] In some embodiments, the target nucleic acid comprises genomic DNA.

[0028] In some embodiments, the sequence information comprises haplotype sequence information.

[0029] Some embodiments of the methods and compositions provided herein include a library of template nucleic acids prepared according to any one of the foregoing methods.

[0030] Some embodiments of the methods and compositions provided herein include a method of obtaining sequence information from a target nucleic acid, said method comprising: (a) compartmentalizing the target nucleic acid into a plurality of first vessels; (b) providing a first index to the target nucleic acid of each first vessel, thereby obtaining a first indexed nucleic acid; (c) combining the first indexed nucleic acids; (d) compartmentalizing the first indexed template nucleic acids into a plurality of second vessels; (e) providing a second index to the first indexed template nucleic of each second vessel, thereby obtaining a second indexed nucleic acid; and (f) obtaining sequence information from the second indexed nucleic acid of each second vessel.

[0031] In some embodiments, step (b) comprises contacting the target nucleic acid with a

plurality of transposomes each comprising a transposase and a transposon sequence comprising the first index under conditions such that at least some of the transposon sequences insert into the target nucleic acid.

[0032] In some embodiments, step (b) comprises contacting the target nucleic acid with a plurality of transposomes each transposon comprising a first transposon sequence comprising a first index, a second transposon sequence noncontiguous with said first transposon sequence, and a transposase associated with the first transposon sequence and the second transposon sequence.

[0033] In some embodiments, step (d) comprises removing the transposase from the compartmentalized first indexed template nucleic acids.

[0034] In some embodiments, the transposase is removed subsequent to step (b).

[0035] In some embodiments, the transposase is removed prior to step (f).

[0036] In some embodiments, removing the transposase comprises a method selected from the group consisting of adding a detergent, changing temperature, changing pH, adding a protease, adding a chaperone, and adding a strand-displacing polymerase.

[0037] In some embodiments, the first transposon sequences comprises a first primer site and the second transposon sequences comprises a second primer site.

[0038] In some embodiments, the first primer site further comprises a first barcode and the second primer site further comprises a second barcode.

[0039] In some embodiments, the first barcode and second barcode are different.

[0040] In some embodiments, step (b) comprises amplifying the target nucleic acid with at least one primer comprising the first index.

[0041] In some embodiments, step (b) comprises ligating the target nucleic acid with at least one primer comprising the first index.

[0042] In some embodiments, the first index provided to the target nucleic acid of each first vessel is different.

[0043] In some embodiments, step (a) comprises providing each first vessel with an amount of target nucleic acid greater than about one or more haploid equivalents of the target nucleic acid.

[0044] In some embodiments, step (d) comprises providing each vessel with an amount of the first indexed template nucleic acids greater than about one or more haploid equivalents of the

target nucleic acid.

[0045] In some embodiments, step (e) comprises amplifying the first indexed template nucleic with at least one primer comprising the second index.

[0046] In some embodiments, step (e) comprises ligating the first indexed template nucleic with at least one primer comprising the second index.

[0047] In some embodiments, the second index provided to the first indexed template nucleic of each second vessel is different.

[0048] In some embodiments, the target nucleic acid comprises an amplified nucleic acid.

[0049] In some embodiments, the target nucleic acid is obtained by enriching a plurality of nucleic acids for a sequence of interest.

[0050] In some embodiments, the target nucleic acid comprises genomic DNA.

[0051] In some embodiments, step (f) further comprises assembling from sequence data a representation of at least a portion of said template nucleic acid from each vessel.

[0052] Some embodiments of the methods and compositions provided herein include a method preparing a library of template nucleic acids to obtain sequence information from a target nucleic acid, said method comprising: (a) compartmentalizing the target nucleic acid into a plurality of first vessels; (b) providing a first index to the target nucleic acid of each first vessel, thereby obtaining a first indexed nucleic acid; (c) combining the first indexed nucleic acids; (d) compartmentalizing the first indexed template nucleic acids into a plurality of second vessels; and (e) providing a second index to the first indexed template nucleic of each second vessel, thereby obtaining a second indexed nucleic acid.

[0053] In some embodiments, step (b) comprises contacting the target nucleic acid with a plurality of transposomes each comprising a transposase and a transposon sequence comprising the first index under conditions such that at least some of the transposon sequences insert into the target nucleic acid.

[0054] In some embodiments, step (b) comprises contacting the target nucleic acid with a plurality of transposomes each transposon comprising a first transposon sequence comprising a first index, a second transposon sequence noncontiguous with said first transposon sequence, and a transposase associated with the first transposon sequence and the second transposon sequence.

[0055] In some embodiments, step (d) comprises removing the transposase from the compartmentalized first indexed template nucleic acids.

[0056] In some embodiments, removing the transposase comprises a method selected from the group consisting of adding a detergent, changing temperature, changing pH, adding a protease, adding a chaperone, and adding a polymerase.

[0057] In some embodiments, the first transposon sequences comprises a first primer site and the second transposon sequences comprises a second primer site.

[0058] In some embodiments, the first primer site further comprises a first barcode and the second primer site further comprises a second barcode.

[0059] In some embodiments, the first barcode and second barcode are different.

[0060] In some embodiments, step (b) comprises amplifying the target nucleic acid with at least one primer comprising the first index.

[0061] In some embodiments, step (b) comprises ligating the target nucleic acid with at least one primer comprising the first index.

[0062] In some embodiments, the first index provided to the target nucleic acid of each first vessel is different.

[0063] In some embodiments, step (a) comprises providing each first vessel with an amount of target nucleic acid greater than about one or more haploid equivalents of the target nucleic acid.

[0064] In some embodiments, step (d) comprises providing each vessel with an amount of the first indexed template nucleic acids greater than about one or more haploid equivalents of the target nucleic acid.

[0065] In some embodiments, step (e) comprises amplifying the first indexed template nucleic with at least one primer comprising the second index.

[0066] In some embodiments, step (e) comprises ligating the first indexed template nucleic with at least one primer comprising the second index.

[0067] In some embodiments, the second index provided to the first indexed template nucleic of each second vessel is different.

[0068] In some embodiments, the target nucleic acid comprises an amplified nucleic acid.

[0069] In some embodiments, the target nucleic acid is obtained by enriching a plurality of nucleic acids for a sequence of interest either before or after transposition.

[0070] In some embodiments, the target nucleic acid comprises genomic DNA.

[0071] Some embodiments of the methods and compositions provided herein include a library of template nucleic acids prepared according to any one of the foregoing methods.

BRIEF DESCRIPTION OF THE DRAWINGS

[0072]

FIG. 1 depicts a schematic of a transposome comprising a dimeric transposase and two non-contiguous transposon sequences, and a transposome comprising a dimeric transposase and a contiguous transposon sequence..

FIG. 2 depicts a method of preparing a transposome with a linker comprising a complementary double-stranded sequence.

FIG. 3 depicts an embodiment of making a template library using transposomes comprising transposon sequences comprising a single stranded linker coupling the two transposon sequences in each transposome in a 5'-5' orientation. Sequences are extended using primers

FIG. 4 depicts a scheme for preparing template nucleic acids to obtain sequence information in which a target nucleic acid is compartmentalized into 96 tubes, indexed by insertion of Tn5 derived transposons, indexed nucleic acids are combined, and further compartmentalized into 96 tubes, further indexed by amplification, the twice indexed nucleic acids may then be combined.

FIG. 5 depicts a schematic embodiment for obtaining haplotype sequence information in which a template nucleic acid is indexed with the barcode of a transposon, and with a primer. A template nucleic acid is prepared by insertion of a looped transposon into a target nucleic acid. The template nucleic acid is diluted into compartments. The template nucleic acid of each compartment is indexed by amplification with a primer. Indexed template nucleic acids are sequenced, aligned, and sequence representation is obtained.

FIG. 6 depicts a scheme that includes preparing a target nucleic acid using matepair and rolling circle amplification, followed by insertion of transposomes into the target nucleic acid, dilution of the target nucleic acid to obtain haplotype information, removal of the transposase by addition of SDS, generation of shotgun libraries, indexing and sequencing.

FIG. 7 depicts a scheme that includes preparing a target nucleic acid using hairpin transposition and rolling circle amplification, followed by insertion of transposomes into the target nucleic acid, dilution of the target nucleic acid, removal of the transposase by addition of SDS, generation of shotgun libraries, indexing and sequencing to obtain haplotype information.

FIG. 8 depicts an example scheme for generation of mate pair libraries.

FIG. 9 depicts an example scheme for generation of mate pair libraries.

FIG. 10 is a graph depicting a model of error rate in sequence information for the number of times a particular sequence associated with a barcode is sequenced.

FIG. 11 depicts images of agarose gels showing oligonucleotides linked with 5'-5' bisoxyamine coupling, in which looped precursor transposons are indicated by the dimer band.

FIG. 12 is an image of an agarose gel showing the apparent molecular weight of a transposed target nucleic acid associated with transposase (left lane), and without transposase (+ 0.1% SDS, middle lane).

FIG. 13 summarizes that haplotype blocks up to 100 kb were observed for samples in which transposase was removed by SDS post-dilution.

FIG. 14 depicts a graph showing the frequencies of sequencing reads for particular distances between neighboring aligned reads for template nucleic acids prepared by adding SDS to remove transposase before dilution to obtain haplotype information, or after dilution to obtain haplotype information.

FIG. 15 shows a graph of barcode indices and proportion of reads and demonstrates that all 9216 different compartments in a 96 X 96 indexing scheme were observed.

FIG. 16 depicts a Pile up analysis of haplotype information obtained using transposomes comprising Mu.

DETAILED DESCRIPTION

[0073] Embodiments of the present invention relate to sequencing nucleic acids. In particular, embodiments of the methods and compositions provided herein relate to preparing nucleic acid templates and obtaining sequence data therefrom. Methods and compositions provided herein are related to the methods and compositions provided in U.S. Patent Application Pub. No. 2012/0208705, U.S. Patent Application Pub. No. 2012/0208724 and Int. Patent Application Pub. No. WO 2012/061832. Some embodiments of the present invention relate to preparing templates to obtain haplotype sequence information from a target nucleic acid, and obtaining haplotype sequence information from such templates. More embodiments relate to preparing templates to obtain sequence information from a strand of a double-stranded target nucleic acid, and obtaining sequence information from such templates. Particular embodiments provided herein relate to the use of integrases, for example transposases, to maintain physical proximity of associated ends of fragmented nucleic acids; and to the use of virtual compartments to enable the use of high concentrations of nucleic acids.

[0074] Obtaining haplotype information from a target nucleic acid includes distinguishing between different alleles (*e.g.,* SNPs, genetic anomalies, etc.) in a target nucleic acid. Such methods are useful to characterize different alleles in a target nucleic acid, and to reduce the

error rate in sequence information. Generally, methods to obtain haplotype sequence information include obtaining sequence information for a portion of a template nucleic acid. In one embodiment, a template nucleic acid can be diluted and sequence information obtained from an amount of template nucleic acid equivalent to about a haplotype of the target nucleic acid.

[0075] In further embodiments, a template nucleic acid can be compartmentalized such that multiple copies of a chromosome can be present in the same compartment, as a result of dual or multiple indexing provided herein, a haplotype can still also be determined. In other words, a template nucleic acid can be prepared using virtual compartments. In such embodiments, a nucleic acid can be distributed between several first compartments, providing a first index to the nucleic acid of each compartment, combining the nucleic acids, distributing the nucleic acid between several second compartments, and providing a second index to the nucleic acid of each compartment. Advantageously, such indexing enables haplotype information to be obtained at higher concentrations of nucleic acid compared to the mere dilution of a nucleic acid in a single compartment to an amount equivalent to a haplotype of the nucleic acid.

[0076] In some embodiments provided herein, template libraries are prepared using transposomes. In some such libraries, the target nucleic acid may be fragmented. Accordingly, some embodiments provided herein relate to methods for maintaining sequence information for the physical contiguity of adjacent fragments. Such methods include the use of integrases to maintain the association of template nucleic acid fragments adjacent in the target nucleic acid. Advantageously, such use of integrases to maintain physical proximity of fragmented nucleic acids increases the likelihood that fragmented nucleic acids from the same original molecule, e,g, chromosome, will occur in the same compartment.

[0077] Other embodiments provided herein relate to obtaining sequence information from each strand of a nucleic acid which can be useful to reduce the error rate in sequencing information. Methods to prepare libraries of template nucleic acids for obtaining sequence information from each strand of a nucleic acid can be prepared such that each strand can be distinguished, and the products of each strand can also be distinguished.

[0078] Some of the methods provided herein include methods of analyzing nucleic acids. Such methods include preparing a library of template nucleic acids of a target nucleic acid, obtaining sequence data from the library of template nucleic acids, and assembling a sequence representation of the target nucleic acid from such sequence data.

[0079] Generally, the methods and compositions provided herein are related to the methods and compositions provided in U.S. Patent Application Pub. No. 2012/0208705, U.S. Patent Application Pub. No. 2012/0208724 and Int. Patent Application Pub. No. WO 2012/061832. The methods provided herein relate to the use of transposomes useful to insert features into a target nucleic acid. Such features including fragmentation sites, primer sites, barcodes, affinity tags, reporter moieties, etc.

[0080] In a method useful with the embodiments provided herein, a library of template nucleic acids is prepared from a target nucleic acid. The library is prepared by inserting a plurality of unique barcodes throughout the target nucleic acid. In some embodiments, 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 sequence can be identified or designated to be paired with one another. The pairing can be informative so that a first barcode is associated with a second barcode. Advantageously, the paired barcode sequences can be used to assemble sequencing data from the library of template nucleic acids. For example, identifying a first template nucleic acid comprising a first barcode sequence and a second template nucleic acid comprising a second barcode sequence that is paired with the first indicates that the first and second template nucleic acids represent sequences adjacent to one another in a sequence representation of the target nucleic acid. Such methods can be used to assemble a sequence representation of a target nucleic acid *de novo,* without the requirement of a reference genome.

Definitions

[0081] As used herein the term "nucleic acid" and/or "oligonucleotide" and/or grammatical equivalents thereof can refer to at least two nucleotide monomers linked together. A nucleic acid can generally contain phosphodiester bonds; however, in some embodiments, nucleic acid analogs may have other types of backbones, comprising, for example, phosphoramide (Beaucage, et al., Tetrahedron, 49:1925 (1993); Letsinger, J. Org. Chem., 35:3800 (1970); Sprinzl, et al., Eur. J. Biochem., 81:579 (1977); Letsinger, et al., Nucl. Acids Res., 14:3487 (1986); Sawai, et al., Chem. Lett., 805 (1984), Letsinger, et al., J. Am. Chem. Soc., 110:4470 (1988); and Pauwels, et al., Chemica Scripta, 26:141 (1986)), phosphorothioate (Mag, et al., Nucleic Acids Res., 19:1437 (1991); and U.S. Pat. No. 5,644,048), phosphorodithioate (Briu, et al., J. Am. Chem. Soc., 111:2321 (1989), O-methylphosphoroamidite linkages (see Eckstein, Oligonucleotides and Analogues: A Practical Approach, Oxford University Press), and peptide nucleic acid backbones and linkages (see Egholm, J. Am. Chem. Soc., 114:1895 (1992); Meier, et al., Chem. Int. Ed. Engl., 31:1008 (1992); Nielsen, Nature, 365:566 (1993); Carlsson, et al., Nature, 380:207 (1996)).

[0082] Other analog nucleic acids include those with positive backbones (Denpcy, et al., Proc. Natl. Acad. Sci. USA, 92:6097 (1995)); non-ionic backbones (U.S. Pat. Nos. 5,386,023; 5,637,684; 5,602,240; 5,216,141; and 4,469,863; Kiedrowshi, et al., Angew. Chem. Intl. Ed. English, 30:423 (1991); Letsinger, et al., J. Am. Chem. Soc., 110:4470 (1988); Letsinger, et al., Nucleosides & Nucleotides, 13:1597 (1994); Chapters 2 and 3, ASC Symposium Series 580, "Carbohydrate Modifications in Antisense Research", Ed. Y. S. Sanghui and P. Dan Cook; Mesmaeker, et al., Bioorganic & Medicinal Chem. Lett., 4:395 (1994); Jeffs, et al., J. Biomolecular NMR, 34:17 (1994); Tetrahedron Lett., 37:743 (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).

[0083] Modifications of the ribose-phosphate backbone may be done to facilitate the addition of additional moieties such as labels, or to increase the stability of such molecules under certain conditions. In addition, mixtures of naturally occurring nucleic acids and analogs can be made. Alternatively, mixtures of different nucleic acid analogs, and mixtures of naturally occurring nucleic acids and analogs may be made. The nucleic acids may be single stranded or double stranded, as specified, or contain portions of both double stranded or single stranded sequence. The nucleic acid may be DNA, for example, genomic or cDNA, RNA or a hybrid, from single cells, multiple cells, or from multiple species, as with metagenomic samples, such as from environmental samples, further from mixed samples for example mixed tissue samples or mixed samples for different individuals of the same species, disease samples such as cancer related nucleic acids, and the like. A nucleic acid can contain any combination of deoxyribo- and ribonucleotides, and any combination of bases, including uracil, adenine, thymine, cytosine, guanine, inosine, xanthanine, hypoxanthanine, isocytosine, isoguanine, and base analogs such as nitropyrrole (including 3-nitropyrrole) and nitroindole (including 5-nitroindole), etc.

[0084] In some embodiments, a nucleic acid can include at least one promiscuous base. Promiscuous bases can base-pair with more than one different type of base. In some embodiments, a promiscuous base can base-pair with at least two different types of bases and no more than three different types of bases. An example of a promiscuous base includes inosine that may pair with adenine, thymine, or cytosine. Other examples include hypoxanthine, 5-nitroindole, acylic 5-nitroindole, 4-nitropyrazole, 4-nitroimidazole and 3-nitropyrrole (Loakes et al., Nucleic Acid Res. 22:4039 (1994); Van Aerschot et al., Nucleic Acid Res. 23:4363 (1995); Nichols et al., Nature 369:492 (1994); Bergstrom et al., Nucleic Acid Res. 25:1935 (1997); Loakes et al., Nucleic Acid Res. 23:2361 (1995); Loakes et al., J. Mol. Biol. 270:426 (1997); and Fotin et al., Nucleic Acid Res. 26:1515 (1998)). Promiscuous bases that can base-pair with at least three, four or more types of bases can also be used.

[0085] 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).

[0086] 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.

[0087] 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. 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.

[0088] 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

[0089] A "transposome" comprises an integration enzyme such as an integrase or transposase, and a nucleic acid comprising an integration recognition site, such as a transposase recognition site. In embodiments provided herein, the transposase can form a functional complex with a transposase recognition site that is capable of catalyzing a transposition reaction. The transposase may bind to the transposase recognition site and insert the transposase recognition site into a target nucleic acid in a process sometimes termed "tagmentation". In some such insertion events, one strand of the transposase recognition site may be transferred into the target nucleic acid. **FIG. 1** depicts two examples of transposomes. In one example, a transposome (10) comprises a dimeric transposase comprising two subunits (20), and two non-contiguous transposon sequences (30). In another example, a transposome (50) comprises a transposase comprises a dimeric transposase comprising two subunits (60), and a contiguous transposon sequence (70).

[0090] Some embodiments can include the use of a hyperactive Tn5 transposase and a Tn5-type transposase recognition site (Goryshin and Reznikoff, J. Biol. Chem., 273:7367 (1998)), or MuA transposase and a Mu transposase recognition site comprising R1 and R2 end sequences (Mizuuchi, K., Cell, 35: 785, 1983; Savilahti, H, et al., EMBO J., 14: 4893, 1995). ME sequences can also be used as optimized by a skilled artisan.

[0091] More examples of transposition systems that can be used with certain embodiments of the compositions and methods provided herein include *Staphylococcus aureus* Tn552 (Colegio et al., J. Bacteriol., 183: 2384-8, 2001; Kirby C et al., Mol. Microbiol., 43: 173-86, 2002), Ty1 (Devine & Boeke, Nucleic Acids Res., 22: 3765-72, 1994 and International Publication WO 95/23875), Transposon Tn7 (Craig, N L, Science. 271: 1512, 1996; Craig, N L, Review in: Curr Top Microbiol Immunol., 204:27-48, 1996), Tn/O and IS10 (Kleckner N, et al., Curr Top Microbiol Immunol., 204:49-82, 1996), Mariner transposase (Lampe D J, et al., EMBO J., 15: 5470-9, 1996), Tc1 (Plasterk R H, Curr. Topics Microbiol. Immunol., 204: 125-43, 1996), P

Element (Gloor, G B, Methods Mol. Biol., 260: 97-114, 2004), Tn3 (Ichikawa & Ohtsubo, J Biol. Chem. 265:18829-32, 1990), bacterial insertion sequences (Ohtsubo & Sekine, Curr. Top. Microbiol. Immunol. 204: 1-26, 1996), retroviruses (Brown, et al., Proc Natl Acad Sci USA, 86:2525-9, 1989), and retrotransposon of yeast (Boeke & Corces, Annu Rev Microbiol. 43:403-34, 1989). More examples include IS5, 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).

[0092] More examples of integrases that may be used with the methods and compositions provided herein include retroviral integrases and integrase recognition sequences for such retroviral integrases, such as integrases from HIV-1, HIV-2, SIV, PFV-1, RSV.

Transposon sequences

[0093] Some embodiments of the compositions and methods provided herein include transposon sequences. In some embodiments, a transposon sequence includes at least one transposase recognition site. In some embodiments, a transposon sequence includes at least one transposase recognition site and at least one barcode. Transposon sequences useful with the methods and compositions provided herein are provided in U.S. Patent Application Pub. No. 2012/0208705, U.S. Patent Application Pub. No. 2012/0208724 and Int. Patent Application Pub. No. WO 2012/061832. In some embodiments, a transposon sequence includes a first transposase recognition site, a second transposase recognition site, and a barcode or barcodes disposed therebetween.

Transposomes with non-contiguous transposon sequences

[0094] Some transposomes provided herein include a transposase comprising two transposon sequences. In some such embodiments, the two transposon sequences are not linked to one another, in other words, the transposon sequences are non-contiguous with one another. Examples of such transposomes are known in the art, *see e.g.,* U.S. Patent Application Pub. No. 2010/0120098. **FIG. 1** depicts an example transposome (10) comprising a dimeric transposase (20) and two transposon sequences (30).

Looped structures

[0095] In some embodiments, a transposome comprises a transposon sequence nucleic acid that binds two transposase subunits to form a "looped complex" or a "looped transposome." In essence, a transposase complex with contiguous transposons. **FIG. 1** depicts an example transposome (50) comprising a dimeric transposase (60) and a transposon sequence (70). Looped complexes can ensure that transposons are inserted into target DNA while maintaining

ordering information of the original target DNA and without fragmenting the target DNA. As will be appreciated, looped structures may insert primers, barcodes, indexes and the like into a target nucleic acid, while maintaining physical connectivity of the target nucleic acid. In some embodiments, the transposon sequence of a looped transposome can include a fragmentation site such that the transposon sequence can be fragmented to create a transposome comprising two transposon sequences. Such transposomes are useful to ensuring that neighboring target DNA fragments, in which the transposons insert, receive code combinations that can be unambiguously assembled at a later stage of the assay.

Barcodes

[0096] 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 generated during transposition, such as identical flanking genomic DNA sequences (g-codes) at the end of formerly juxtaposed DNA fragments. 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.

[0097] In some embodiments, a transposon sequence comprises at least one barcode. In some embodiments, such as transposomes comprising two non-contiguous transposon sequences, the first transposon sequence comprises a first barcode, and the second transposon sequence comprises a second barcode. In some embodiments, such as in looped transposomes, a transposon sequence comprises a barcode comprising a first barcode sequence and a second barcode sequence. In some of the foregoing embodiments, the first barcode sequence can be identified or designated to be paired with the second barcode sequence. For example, a known first barcode sequence can be known to be paired with a known second barcode sequence using a reference table comprising a plurality of first and second bar code sequences known to be paired to one another.

[0098] 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. The first and second barcode sequences may comprise a bi-code.

[0099] In some embodiments of compositions and methods described herein, barcodes are used in the preparation of template nucleic acids. As will be understood, 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. For example, uniquely identified molecules can be applied to identify individual nucleic acid molecules, in samples having multiple chromosomes, in genomes, in cells, in cell types, in cell disease states, and in species, for example, in haplotype sequencing, in parental allele discrimination, in metagenomic sequencing, and in sample sequencing of a genome.

Linkers

[0100] Some embodiments comprising looped transposomes where a transposase is complexed with contiguous transposons 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, 200, 300, 400, 500 or more nucleotides.

[0101] In some embodiments, a linker can be amplifiable for example by PCR, rolling circle amplification, strand displacement amplification, and the like. In other embodiments, a linker can comprise non-amplifiable moieties. Examples of non-amplifiable linkers include organic chemical 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.

[0102] In some embodiments, the linker comprises a single-stranded nucleic acid. In some embodiments, the linker couples transposon sequences in a 5'-3' orientation, a 5'-5' orientation, or a 3'-3' orientation.

Fragmentation sites

[0103] In some embodiments comprising looped transposomes 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 comprise a restriction endonuclease

site; at least one ribonucleotide cleavable with an RNAse; nucleotide analogues cleavable in the presence of certain chemical agent; a diol linkage cleavable by treatment with periodate; a disulphide group cleavable with a chemical reducing agent; a cleavable moiety that may be subject to photochemical cleavage; and a peptide cleavable by a peptidase enzyme or other suitable means. *See e.g.,* U.S. Patent Application Pub. No. 2012/0208705, U.S. Patent Application Pub. No. 2012/0208724 and Int. Patent Application Pub. No. WO 2012/061832.

Primer sites

[0104] In some embodiments, a transposon sequence can include 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 transposon sequence can include at least a first primer site useful for amplification, sequencing, and the like. In some embodiments comprising looped transposomes, a linker can include a sequencing adaptor. In more embodiments comprising looped transposomes, 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.

[0105] 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. 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.

Affinity tags

[0106] In some embodiments, a transposon sequence or transposase can include an affinity tag. In some embodiments comprising looped transposomes a linker can comprise an affinity tag. Affinity tags can be useful for a variety of applications, for example the bulk separation of target nucleic acids hybridized to hybridization tags. Additional application include, but are not limited to, using affinity tags for purifying transposase/transposon complexes and transposon inserted target DNA, for example. As used herein, the term "affinity tag" and grammatical equivalents can refer to a component of a multi-component complex, wherein the components of the multi-component complex specifically interact with or bind to each other. For example an

affinity tag can include biotin or poly-His that can bind streptavidin or nickel, respectively. Other examples of multiple-component affinity tag complexes are listed, for example, U.S. Patent Application Pub. No. 2012/0208705, U.S. Patent Application Pub. No. 2012/0208724 and Int. Patent Application Pub. No. WO 2012/061832.

Reporter moieties

[0107] In some embodiments of the compositions and methods described herein, a transposon sequence or transposase can include a reporter moiety. In some embodiments comprising looped transposomes 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 a signal includes, but is not limited to, a fluorescent, a chemiluminescent, a bioluminescent, a phosphorescent, a radioactive, a calorimetric, an ion activity, an electronic or an electrochemiluminescent signals. Example reporter moieties are listed, for example, U.S. Patent Application Pub. No. 2012/0208705, U.S. Patent Application Pub. No. 2012/0208724 and Int. Patent Application Pub. No. WO 2012/061832.

Certain methods of making transposon sequences

[0108] The transposon sequences provided herein can be prepared by a variety of methods. Exemplary methods include direct synthesis , hairpin extension methods, and PCR. In some embodiments, transposon sequences may be prepared by direct synthesis. For example, a transposon sequence comprising a nucleic acid may be prepared by methods comprising chemical synthesis. Such methods are well known in the art, *e.g.,* solid phase synthesis using phosphoramidite precursors such as those derived from protected 2'-deoxynucleosides, ribonucleosides, or nucleoside analogues. Example methods of preparing transposon sequencing can be found in, for example, U.S. Patent Application Pub. No. 2012/0208705, U.S. Patent Application Pub. No. 2012/0208724 and Int. Patent Application Pub. No. WO 2012/061832.

[0109] In some embodiments comprising looped transposomes transposon sequences comprising a single stranded linker can be prepared. In some embodiments, the linker couples the transposon sequences of a transposome such that a transposon sequence comprising a first transposase recognition sequence is coupled to a second transposon sequence comprising a second transposase recognition sequence in a 5' to 3' orientation. In some embodiments, the linker couples a transposon sequence comprising a first transposase recognition sequence to a second transposon sequence comprising a second transposase recognition sequence in a 5' to 5' orientation or in a 3' to 3' orientation. Coupling transposon

sequences of a transposome in either a 5' to 5' orientation or in a 3' to 3' orientation can be advantageous to prevent transposase recognition elements, in particular mosaic elements (ME or M), from interacting with one another. For example, coupled transposon sequences can be prepared by preparing transposon sequences comprising either an aldehyde group or oxyamine group. The aldehyde and oxyamine groups can interact to form a covalent bond thus coupling the transposon sequences.

[0110] In some embodiments, transposomes comprising complementary sequences can be prepared. **FIG. 2** illustrates an embodiment in which a transposase is loaded with transposon sequences comprising complementary tails. The tails hybridize to form a linked transposon sequence. Hybridization may occur in dilute conditions to decrease the likelihood of hybridization between transposomes.

Target nucleic acids

[0111] 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, mixed samples of nucleic acids, polyploidy DNA (i.e., plant DNA), mixtures thereof, and hybrids thereof. In a preferred embodiment, genomic DNA or amplified copies thereof are used as the target nucleic acid. In another preferred embodiment, cDNA, mitochondrial DNA or chloroplast DNA is used.

[0112] 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 or 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 or 20 times or more.

[0113] 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

[0114] 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); archeaon, such as crenarchaeota, nanoarchaeota or euryarchaeotia; or eukaryotic such as fungi, (for example, yeasts), plants, protozoans and other parasites, and animals (including insects (for example, *Drosophila* spp.), nematodes (*e.g., Caenorhabditis elegans*), and mammals (for example, rat, mouse, monkey, non-human primate and human). Target nucleic acids and template nucleic acids can be enriched for certain sequences of interest using various methods well known in the art. Examples of such methods are provided in Int. Pub. No. WO/2012/108864. In some embodiments, nucleic acids may be further enriched during methods of preparing template libraries. For example, nucleic acids may be enriched for certain sequences, before insertion of transposomes after insertion of transposomes and/or after amplification of nucleic acids.

[0115] In addition, in some embodiments, target nucleic acids and/or template nucleic acids can be highly purified, for example, nucleic acids can be at least about 70%, 80%, 90%, 95%, 96%, 97%, 98%, 99%, or 100% free from contaminants before use with the methods provided herein. In some embodiments, it is beneficial to use methods known in the art that maintain the quality and size of the target nucleic acid, for example isolation and/or direct transposition of target DNA may be performed using agarose plugs. Transposition can also be performed directly in cells, with population of cells, lysates, and non-purified DNA.

Certain methods of preparing template nucleic acids

[0116] Some embodiments include methods of preparing template nucleic acids. As used herein, "template nucleic acid" can refer to a substrate for obtaining sequence information. In some embodiments, a template nucleic acid can include a target nucleic acid, a fragment thereof, or any copy thereof comprising at least one transposon sequence, a fragment thereof, or any copy thereof. In some embodiments, a template nucleic acid can include a target nucleic acid comprising a sequencing adaptor, such as a sequencing primer site.

[0117] 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 or sequences into the target nucleic acid.

[0118] 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.

[0119] 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. As used herein, the term "host tag" or "g-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.

[0120] 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.

[0121] Some embodiments include selecting conditions sufficient to achieve insertion of a transposon sequence or sequences into a target nucleic acid, but not into another transposon sequence or sequences. A variety of methods can be used to reduce the likelihood that a transposon sequence inserts into another transposon sequence. Examples of such methods useful with the embodiments provided herein can be found in for example, U.S. Patent Application Pub. No. 2012/0208705, U.S. Patent Application Pub. No. 2012/0208724 and Int. Patent Application Pub. No. WO 2012/061832.

[0122] In some embodiments, conditions may be selected so that the average distance in a target nucleic acid between integrated transposon sequences is at least about 5, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, or more consecutive nucleotides. In some embodiments, the average distance in a target nucleic acid between integrated transposon sequences is at least about 100, 200, 300, 400, 500, 600, 700, 800, 900, 1000, or more consecutive nucleotides. In some embodiments, the average distance in a target nucleic acid between integrated

transposon sequences is at least about 1 kb, 2 kb, 3 kb, 4 kb, 5 kb, 6 kb, 7 kb, 8 kb, 90 kb, 100 kb, or more consecutive nucleotides. In some embodiments, the average distance in a target nucleic acid between integrated transposon sequences is at least about 100 kb, 200 kb, 300 kb, 400 kb, 500 kb, 600 kb, 700 kb, 800 kb, 900 kb, 1000 kb, or more consecutive nucleotides. As will be understood, some conditions that may be selected include contacting a target nucleic acid with a certain number of transposon sequences.

[0123] Some embodiments of the methods described herein include selecting conditions sufficient to achieve at least a portion of transposon sequences integrated into a target nucleic acid that are different. In preferred embodiments of the methods and compositions described herein, 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 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. In some embodiments, at least a portion of the transposon sequences integrated into a target nucleic acid are the same.

[0124] 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. In more embodiments, adaptors comprising primer sites can be ligated to each end of a nucleic acid, and the nucleic amplified from such primer sites.

[0125] 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.

[0126] Some methods of preparing template nucleic acids include inserting transposon sequences comprising single-stranded linkers. **FIG. 3** illustrates an example in which transposon sequences (ME-P1-linker-P2-ME; mosaic end-primer site 1-linker-primer site 2-mosaic end) are inserted into a target nucleic acid. The target nucleic acid having the inserted transposon/linker sequences can be extended and amplified.

[0127] In one embodiment of the compositions and methods described herein, transposomes are used that have symmetrical transposable end sequences to produce an end-tagged target nucleic acid fragment (tagmented fragment or tagment). Each tagmented fragment therefore contains identical ends, lacking directionality. A single primer PCR, using the transposon end sequences, can then be employed to amplify the template copy number from $2n$ to $2n*2^x$ where x corresponds to the number of PCR cycles. In a subsequent step, PCR with primers can add additional sequences, such as sequencing adapter sequences.

[0128] 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 use in 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. Universal primer sites can be introduced into nucleic acids by various methods well known in the art, for example, ligation of primer sites to nucleic acids, amplification of nucleic acids using tailed primers, and insertion of a transposon sequence comprising a universal primer site.

Targeted insertion

[0129] In some embodiments of the methods and compositions provided herein, transposon sequences may be inserted at particular targeted sequences of a target nucleic acid. Transposition into dsDNA can be more efficient than into ssDNA targets. In some embodiments, dsDNA is denatured into ssDNA and annealed with oligonucleotide probes (20-200 bases). These probes create sites of dsDNA that can be efficiently used as integration sites with transposomes provided herein. In some embodiments, dsDNA can be targeted using D-loop formation with recA-coated oligo probes, and subsequent triplex formation. In some such embodiments, the replication fork structure is the preferred substrate for transposomes comprising Tn4430 transposase. In more embodiments, 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.

[0130] In some embodiments, transposomes comprising a transposase having a preferred substrate of mismatched positions in a target nucleic acid may be used to target insertion into the target nucleic acid. For example, some MuA transposases, such as HYPERMU (Epicenter), have a preference for mismatched targets. In some such embodiments, oligonucleotide probes comprising a mismatch are annealed to a single-stranded target nucleic acid. Transposomes comprising MuA transposases, such as HYPERMU, can be used to target the mismatched sequences of the target nucleic acid.

Fragmenting template nucleic acids

[0131] Some embodiments of preparing a template nucleic acid can include fragmenting a target nucleic acid. In some embodiments, insertion of transposomes comprising non-contiguous transposon sequences can result in fragmentation of a target nucleic acid. In some embodiments comprising looped transposomes a target nucleic acid comprising transposon sequences can be fragmented at the fragmentation sites of the transposon sequences. Further examples of method useful to fragment target nucleic acids useful with the embodiments provided herein can be found in for example, U.S. Patent Application Pub. No. 2012/0208705, U.S. Patent Application Pub. No. 2012/0208724 and Int. Patent Application Pub. No. WO 2012/061832.

Tagging single molecules

[0132] The present disclosure provides methods for tagging molecules so that individual molecules can be tracked and identified. The bulk data can then be deconvoluted and converted back to the individual molecule. The ability to distinguish individual molecules and relate the information back to the molecule of origin is especially important when processes from original molecule to final product change the (stoichiometric) representation of the original population. For example, amplification leads to duplication (*e.g.,* PCR duplicates or biased amplification) that can skew the original representation. This can alter the methylation state call, copy number, allelic ratio due to non-uniform amplification and/or amplification bias. By identifying individual molecules, code-tagging distinguishes between identical molecules after processing. As such, duplications, and amplification bias can be filtered out, allowing accurate determination of the original representation of a molecule or population of molecules.

[0133] An advantage of uniquely tagging single molecules is that identical molecules in the original pool become uniquely identified by virtue of their tagging. In further downstream analyses, these uniquely tagged molecules can now be distinguished. This technique can be exploited in assay schemes in which amplification is employed. For example, amplification is known to distort the original representation of a mixed population of molecules. If unique tagging were not employed, the original representation (such as copy number or allelic ratio) would need to account for the biases (known or unknown) for each molecule in the

representation. With unique tagging, the representation can accurately be determined by removing duplicates and counting the original representation of molecules, each having a unique tag. Thus, cDNAs can be amplified and sequenced, without fear of bias because the data can be filtered so that only authentic sequences or sequences of interest are selected for further analysis. Accurate reads can be constructed by taking the consensus across many reads with the same barcode.

[0134] In some embodiments of the compositions and methods described herein, it is preferred to tag the original population in the early stages of the assay, although tagging can occur at later stages if the earlier steps do not introduce bias or are not important. In any of these applications, the complexity of the barcode sequences should be larger than the number of individual molecules to be tagged. This ensures that different target molecules receive different and unique tags. As such, a pool of random oligonucleotides of a certain length (e.g., 5, 10, 20, 30, 40, 50, 100 or 200 nucleotides in length) is desirable. A random pool of tags represents a large complexity of tags with code space 4" 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.

[0135] In one embodiment of the compositions and methods described herein, individual molecules (such as target DNA) are attached to unique labels, such as unique oligo sequences and/or barcodes. Attachment of the labels can occur through ligation, coupling chemistry, adsorption, insertion of transposon sequences, etc. Other means include amplification (such as by PCR, RCA or LCR), copying (such as addition by a polymerase), and non-covalent interactions.

[0136] Specific methods comprise including barcodes (*e.g.,* designed or random sequences) to PCR primers so that each template will receive an individual code within the code space, thereby yielding unique amplicons that can be discriminated from other amplicons. This concept can be applied to any method that uses polymerase amplification, such as GoldenGate assays as disclosed in U.S. Patent No. 7,582,420, No. 7,955,794, and No. 8,003,354. Code-tagged target sequences can be circularized and amplified by methods such as rolling-circle amplification to yield code-tagged amplicons. Similarly, the code can also be added to RNA

Methods of analyzing template nucleic acids

[0137] Some embodiments of the technology described herein include methods of analyzing template nucleic acids. In such embodiments, sequencing information can be obtained from template nucleic acids and this information can be used to generate a sequence representation of one or more target nucleic acids.

[0138] In some embodiments of the sequencing methods described herein, a linked read strategy may be used. A linked read strategy can include identifying sequencing data that links

at least two sequencing reads. For example, a first sequencing read may contain a first marker, and a second sequencing read may contain a second marker. The first and second markers can identify the sequencing data from each sequencing read to be adjacent in a sequence representation of the target nucleic acid. In some embodiments of the compositions and methods described herein, 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 other embodiments, markers can comprise a first host tag and a second host tag. In more embodiments, markers can comprise a first barcode sequence with a first host tag, and a second barcode sequence with a second host tag.

[0139] An exemplary embodiment of a method for sequencing a template nucleic acid can comprise the following steps: (a) sequence the first barcode sequence using a sequencing primer hybridizing to the first primer site; and (b) sequence the second barcode sequence using a sequencing primer hybridizing to the second primer. The result is two sequence reads that help link the template nucleic acid to its genomic neighbors. Given long enough reads, and short enough library fragments, these two reads can be merged informatically to make one long read that covers the entire fragment. Using the barcode sequence reads and the 9 nucleotide duplicated sequence present from the insertion, reads can now be linked to their genomic neighbors to form much longer "linked reads" *in silic*o.

[0140] 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.

[0141] In some embodiments of the sequencing technology described herein, sequence analysis is performed in real time. For example, real time sequencing can be performed by simultaneously acquiring and analyzing sequencing data. 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 No. WO 2010/062913.

[0142] In an exemplary embodiment of a method for assembling short sequencing reads using a linked read strategy, transposon sequences comprising barcodes are inserted into genomic DNA, a library is prepared and sequencing data is obtained for the library of template nucleic acids. Blocks of templates can be assembled by identifying paired barcodes and then larger contigs are assembled. In one embodiment, the assembled reads can be further assembled into larger contigs through code pairing using overlapping reads.

[0143] Some embodiments of the sequencing technology described herein include error detection and correction features. Examples of errors can include errors in base calls during a sequencing process, and errors in assembling fragments into larger contigs. As would be

understood, error detection can include detecting the presence or likelihood of errors in a data set, and as such, detecting the location of an error or number of errors may not be required. For error correction, information regarding the location of an error and/or the number of errors in a data set is useful. Methods for error correction are well known in the art. Examples include the use of hamming distances, and the use of a checksum algorithm (*See, e.g.*,U.S. Patent Application Publication No. 2010/0323348; U.S. Patent No. 7,574,305; and U.S. Patent No. 6,654,696).

Nested libraries

[0144] An alternative method involves the junction tagging methods above and preparation of nested sequencing libraries. 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.

[0145] 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 random methods 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 of the previously-described method, code-tagged fragments can be preferentially isolated by using transposons that contain a biotin or other affinity functionality for downstream enrichment purposes. Subsequent library preparation converts the nested DNA fragments into sequencing templates. Paired-end sequencing results in determination of the sequence 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.

Sequencing methods

[0146] The methods and composition 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.

[0147] Some embodiments of the sequencing methods described herein include sequencing by synthesis (SBS) technologies, for example, pyrosequencing techniques. Pyrosequencing detects the release of inorganic pyrophosphate (PP_i) as particular nucleotides are incorporated into the nascent strand (Ronaghi et al., Analytical Biochemistry 242(1): 84-9 (1996); Ronaghi,

M. Genome Res. 11(1):3-11 (2001); Ronaghi et al., Science 281(5375):363 (1998); U.S. Patent No. 6,210,891; U.S. Patent No. 6,258,568 and U.S. Patent No. 6,274,320).

[0148] In another example type of SBS, cycle sequencing is accomplished by stepwise addition of reversible terminator nucleotides containing, for example, a cleavable or photobleachable dye label as described, for example, in U.S. Patent No. 7,427,67, U.S. Patent No. 7,414,1163 and U.S. Patent No. 7,057,026. This approach, which is being commercialized by Illumina Inc., is also described in International Patent Application Publication Nos. WO 91/06678 and WO 07/123744. 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.

[0149] Additional exemplary SBS systems and methods which can be utilized with the methods and compositions 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.

[0150] Some embodiments of the sequencing technology described herein can utilize sequencing by ligation techniques. Such techniques utilize DNA ligase to incorporate nucleotides and identify the incorporation of such nucleotides. Exemplary SBS systems and methods which can be utilized with the compositions and methods 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.

[0151] Some embodiments of the sequencing technology described herein can include techniques such as next-next technologies. One example can include nanopore sequencing techniques (Deamer, D.W. & Akeson, M. "Nanopores and nucleic acids: prospects for ultrarapid sequencing." Trends Biotechnol. 18, 147-151 (2000); Deamer, D. and D. Branton, "Characterization of nucleic acids by nanopore analysis". Acc. Chem. Res. 35:817-825 (2002); Li et al., "DNA molecules and configurations in a solid-state nanopore microscope" Nat. Mater. 2:611-615 (2003)). In such embodiments, the target nucleic acid passes through a nanopore. The nanopore can be a synthetic pore or biological membrane protein, such as α-hemolysin. As the target nucleic acid passes through the nanopore, each base-pair can be identified by measuring fluctuations in the electrical conductance of the pore. (U.S. Patent No. 7,001,792; Soni & Meller, "A. Progress toward ultrafast DNA sequencing using solid-state nanopores." Clin. Chem. 53, 1996-2001 (2007); Healy, K. "Nanopore-based single-molecule DNA analysis." Nanomed. 2:459-481 (2007); Cockroft et al., "A single-molecule nanopore device detects DNA polymerase activity with single-nucleotide resolution." J. Am. Chem. Soc. 130:818-820 (2008)). In some such embodiments, nanopore sequencing techniques can be useful to confirm sequence information generated by the methods described herein.

[0152] Some embodiments of the sequencing technology described herein 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); Korlach, J. et al. "Selective aluminum passivation for targeted immobilization of single DNA polymerase molecules in zero-mode waveguide nanostructures." Proc. Natl. Acad. Sci. USA 105, 1176-1181 (2008)). In one example, single molecule, real-time (SMRT) DNA sequencing technology provided by Pacific Biosciences Inc. can be utilized with the methods described herein. In some embodiments, a SMRT chip or the like may be utilized (e.g., U.S. Patent No. 7,181,122, U.S. Patent No. 7,302,146 and U.S. Patent No. 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.

[0153] SMRT chips and similar technology can be used in association with nucleotide monomers fluorescently labeled on the terminal phosphate of the nucleotide (Korlach J. et al., "Long, processive enzymatic DNA synthesis using 100% dye-labeled terminal phosphate-linked nucleotides." Nucleosides, Nucleotides and Nucleic Acids, 27:1072-1083, 2008). The label is cleaved from the nucleotide monomer on incorporation of the nucleotide into the polynucleotide. Accordingly, the label is not incorporated into the polynucleotide, increasing the signal: background ratio. Moreover, the need for conditions to cleave a label from labeled nucleotide monomers is reduced.

[0154] 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 complementary to the target nucleic acid. In one embodiment, fluorescently-

labeled nucleotide monomers, namely, A, C, G, or T, are delivered one at a time to the target nucleic acid in the presence DNA polymerase. Incorporation of a labeled nucleotide into the polynucleotide complementary to the target nucleic acid is detected, and the position of the fluorescent signal on the glass cover slip indicates the molecule that has been extended. The fluorescent label is removed before the next nucleotide is added to continue the sequencing cycle. Tracking nucleotide incorporation in each polynucleotide strand can provide sequence information for each individual target nucleic acid.

[0155] An additional example of a sequencing platform that can be used in association with the methods described herein is provided by Complete Genomics Inc. Libraries of target nucleic acids can be prepared where target nucleic acid sequences are interspersed approximately every 20 bp with adaptor sequences. The target nucleic acids can be amplified using rolling circle replication, and the amplified target nucleic acids can be used to prepare an array of target nucleic acids. Methods of sequencing such arrays include sequencing by ligation, in particular, sequencing by combinatorial probe-anchor ligation (cPAL).

[0156] In some embodiments using cPAL, about 10 contiguous bases adjacent to an adaptor may be determined. A pool of probes that includes four distinct labels for each base (A, C, T, G) is used to read the positions adjacent to each adaptor. A separate pool is used to read each position. A pool of probes and an anchor specific to a particular adaptor is delivered to the target nucleic acid in the presence of ligase. The anchor hybridizes to the adaptor, and a probe hybridizes to the target nucleic acid adjacent to the adaptor. The anchor and probe are ligated to one another. The hybridization is detected and the anchor-probe complex is removed. A different anchor and pool of probes is then delivered to the target nucleic acid in the presence of ligase.

[0157] 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.

[0158] The methods set forth herein can use arrays having features at any of a variety of densities including, for example, at least about 10 features/cm^2, 100 features/cm^2, 500 features/cm^2, 1,000 features/cm^2, 5,000 features/cm^2, 10,000 features/cm^2, 50,000

features/cm^2, 100,000 features/cm^2, 1,000,000 features/cm^2, 5,000,000 features/cm^2, 10 features/cm^2, 5x 10^7 features/cm^2, 10^8 features/cm^2, 5x 10^8 features/cm^2, 10^9 features/cm^2, 5x 10^9 features/cm^2, or higher.

Surfaces

[0159] 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.

[0160] 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 well known in the art can be used to attach, anchor or immobilize nucleic acids to the surface of the substrate.

Methods for reducing error rates in sequencing data

[0161] Some embodiments of the methods and compositions provided herein include reducing the error rates in sequencing data. In some such embodiments, the sense and antisense strands of a double-stranded target nucleic acid are each associated with a different barcode. Each strand is amplified, sequence information is obtained from multiple copies of the amplified strands, and a consensus sequence representation of the target nucleic acid is generated from the redundant sequence information. Thus, sequence information can originate and be identified from each strand. Accordingly, sequence errors can be identified and reduced where sequence information originating from one strand is inconsistent with sequence information from the other strand.

[0162] In some embodiments, the sense and antisense strands of a target nucleic acid are associated with a different barcode. The barcodes may be associated with the target nucleic acid by a variety of methods including ligation of adaptors and insertion of transposon sequences. In some such embodiments, a Y-adaptor may be ligated to at least one end of a target nucleic acid. The Y-adaptor can include a double-stranded sequence, and non-

complementary strands, each strand comprising a different barcode. The target nucleic acid with ligated Y-adaptor can be amplified and sequenced such that each barcode can be used to identify the original sense or antisense strands. A similar method is described in Kinde I. et al., (2011) PNAS 108:9530-9535. In some embodiments, the sense and antisense strands of a target nucleic acid are associated with a different barcode by inserting transposon sequences provided herein. In some such embodiments, the transposon sequences can comprise non-complementary barcodes.

[0163] Some embodiments of such methods include obtaining sequence information from a strand of a target double-stranded nucleic acid comprising (a) obtaining sequence data from a template nucleic acid comprising a first sequencing adapter and a second sequencing adapter having at least a portion of the double-stranded target nucleic acid disposed therebetween, wherein: (i) the first sequencing adapter comprises a double-stranded first barcode, a single-stranded first primer site and a single-stranded second primer site, wherein the first and second primer sites are non-complementary, and (ii) the second sequencing adapter comprising a double-stranded second barcode, a single-stranded third primer site and a single-stranded fourth primer site, wherein the third and fourth primer sites are non-complementary. In some embodiments, the first primer site of the sense strand of the template nucleic acid and the third primer site of the antisense sense strand of the template nucleic acid comprise the same sequence. In some embodiments, each barcode is different. In some embodiments, the first sequencing adapter comprises a single-stranded hairpin coupling the first primer site and second primer site.

[0164] In another embodiment, each end of a target nucleic acid is associated with an adaptor comprising a different barcode such that extension products from the sense and antisense strand of a nucleic acid can be distinguished from each other. In some embodiments, primer site sequences and barcodes are selected such that extension from a primer annealed to the sense strand yields products that can be distinguished from products of extension from a primer annealed to the antisense strand. In an example, the 3' sense primer site is the same as the 3' antisense primer site, but different from both the 5' sense and 5' antisense primer sites. Extension of primers annealed to the 3' sense primer site and the 3' antisense primer site would yield the following products from each strand:

Sense strand: (5') barcode 2 - [target sequence] - barcode 1 (3')

Antisense strand: (5') barcode 1 - [target sequence] - barcode 2 (3')

[0165] Thus, extension products from the sense and antisense strand of a nucleic acid can be distinguished from each other. An exemplary method is illustrated in Schmitt M.W., et al., PNAS (2012) 109:14508-13. In some such methods, the barcodes and primers sites may be associated with the target nucleic acid by a variety of methods including ligation of adaptors and insertion of transposon sequences. In some embodiments, transposon sequences can be designed to provide adaptors with hairpins. Hairpins provide the ability to maintain the physical

contiguity of the sense and antisense strands of a target nucleic acid. A template nucleic acid can be prepared comprising hairpins using transposon sequences comprising linkers described herein. Examples of linkers include single-stranded nucleic acids.

[0166] Some embodiments of preparing a library of template nucleic acids for obtaining sequence information from each strand of a double-stranded target nucleic acid include (a) providing a population of transposomes comprising a transposase and a first transposon sequence comprising: (i) a first transposase recognition site, a first primer site, and a first barcode, and (ii) a second transposon sequence comprising a second transposase recognition site, a second primer site, and a second barcode, wherein the first transposon sequence is non-contiguous with the second transposon sequence; and (b) contacting the transposomes with a double-stranded nucleic acid under conditions such that said first and second transposon sequences insert into the double-stranded target nucleic acid, thereby preparing a library of template nucleic acids for obtaining sequence information from each strand of the double-stranded target nucleic acid. In some embodiments, the population of transposomes further comprises transposomes comprising a transposase and a transposon sequence comprising a third transposase recognition site and a fourth transposase recognition site having a barcode sequence disposed therebetween, said barcode sequence comprising a third barcode and a fourth barcode having a sequencing adapter disposed therebetween, said sequencing adapter comprising a third primer site and a fourth primer site having a linker disposed therebetween. In some embodiments, the first primer site of the sense strand of the template nucleic acid and the third primer site of the antisense sense strand of the template nucleic acid comprise the same sequence. Some embodiments also include a step (c) selecting for template nucleic acids comprising transposon sequences wherein the first transposon sequence is non-contiguous with the second transposon sequence and transposon sequences comprising a linker. In some embodiments, the linker comprises an affinity tag adapted to bind with a capture probe. In some embodiments, the affinity tag is selected from the group consisting of His, biotin, and streptavidin. In some embodiments, each barcode is different. In some embodiments, the linker comprises a single-stranded nucleic acid. In some embodiments, the target nucleic acid comprises genomic DNA.

Methods for obtaining haplotype information

[0167] Some embodiments of the methods and compositions provided herein include methods of obtaining haplotype information from a target nucleic acid. Haplotype information can include determining the presence or absence of different sequences at specified loci in a target nucleic acid, such as a genome. For example, sequence information can be obtained for maternal and paternal copies of an allele. In a polyploidy organism, sequence information can be obtained for at least one haplotype. Such methods are also useful in reducing the error rate in obtaining sequence information from target nucleic acid.

[0168] Generally, methods to obtain haplotype information include distributing a nucleic acid into one or more compartments such that each compartment comprises an amount of nucleic

acid equivalent to about a haploid equivalent of the nucleic acid, or equivalent to less than about a haploid equivalent of the nucleic acid. Sequence information can then be obtained from each compartment, thereby obtaining haplotype information. Distributing the template nucleic acid into a plurality of vessels increases the probability that a single vessel includes a single copy of an allele or SNP, or that consensus sequence information obtained from a single vessel reflects the sequence information of an allele or SNP. As will be understood, in some such embodiments, a template nucleic acid may be diluted prior to compartmentalizing the template nucleic acid into a plurality of vessels. For example, each vessel can contain an amount of target nucleic acids equal to about a haploid equivalent of the target nucleic acid. In some embodiments, a vessel can include less than about one haploid equivalent of a target nucleic acid.

Method of haplotyping with virtual compartments

[0169] Some methods of obtaining haplotype information provided herein include the use of virtual compartments. Advantageously, some such methods enable compartments to include amounts of nucleic acids equivalent to at least one or more haploid equivalents. In other words, such methods enable the use of higher concentrations of nucleic acids in compartments compared to other methods of haplotyping, thereby increasing the efficiency and yields of various manipulations.

[0170] In some methods to obtain haplotype information with virtual compartments, a nucleic acid is compartmentalized into a plurality of first vessels, and the nucleic acids of each compartment are provided with a first index; the first-indexed nucleic acids are combined, and then compartmentalized into a plurality of second vessels, and the nucleic acids of each compartment are provided with a second index. A template nucleic acid can be prepared by undergoing at least 2, 3, 4, 5, 6, 7, 8, 9, 10 or more rounds of compartmentalizing, indexing, and pooling. In such a manner, a template nucleic acid is provided with a plurality of different indices in a stepwise method. Subsequent to indexing, the indexed template nucleic acids can be pooled and distributed into a plurality of compartments such that each compartment is likely to include an amount of a particular template nucleic acid having a particular combination of indexes that is equivalent to about a haploid equivalent of the target nucleic acid, or equivalent to less than about a haploid equivalent of the target nucleic acid, or equivalent to more than about a haploid equivalent. In other words, each vessel can receive an amount of template nucleic acid comprising more than the equivalent of a haploid equivalent, however, each copy of an allele or SNP is likely to be associated with a different combination of indexes. Accordingly, the number of vessels to compartmentalize a template nucleic acid such that each vessel includes about an amount of template nucleic acid equivalent to a haploid or less of a target nucleic acid can be reduced. In addition, the amount of nucleic acid in each vessel can be greater than the amount of about a haploid equivalent, thereby increasing the efficiency and yields of various manipulations.

[0171] There are various methods to index nucleic acids. For example, in some embodiments,

indexes may be inserted into nucleic acids using transposomes provide herein; indexes can be ligated to nucleic acids; and indexes can be added to nucleic acids during copying, *e.g.*, amplification of a nucleic acid. In some embodiments, a template nucleic acid comprising an index can be prepared using transposomes comprising a contiguous transposon sequence. *See e.g.,* transposome (50) in **FIG. 1.** Insertion of contiguous transposon sequences can result in the preservation of positional information for a particular nucleic acid molecule after distribution of template nucleic acids between several compartments. In some embodiments, a template nucleic acid comprising an index can be prepared using transposomes comprising non-contiguous transposon sequences. *See e.g.,* transposome (10) in **FIG. 1.** Examples of such transposon sequences are set forth in U.S. Patent Application Publication No. 2010/0120098. Insertion of non-contiguous transposon sequences can result in the fragmentation of a particular nucleic acid molecule. Thus, in some embodiments, insertion of non-contiguous transposon sequences into a template nucleic acid can reduce positional information for a particular nucleic acid molecule after distribution of the template nucleic acids between several compartments. In other words, different fragments of a particular nucleic acid molecule can be distributed into different vessels.

[0172] In an example with a diploid genome, after pooling and dilution in compartments, a greater amount of nucleic acids can be added to each compartment since the chance of a copy from the father and copy from the mother of the same region with the same indexes is lower. For example, one copy of the father and one copy of the mother for the same region can be present in the same compartment as long as each contains a different index, for example, one comes from a transposition reaction with a first index (index-1) and the other comes from a transposition reaction with a different first index (index-2). In other words, copies of the same region/chromosome can be present in the same compartment since these can be distinguished by their unique index incorporated in the first transposition reaction. This allows more DNA to be distributed into each compartment compared to alternative dilution methods. The dual indexing scheme creates a total number of virtual compartments of number of initial indexed transposition reactions multiplied by the number of indexed PCR reactions.

[0173] FIG. 4 depicts an example embodiment of obtaining haplotype information using virtual compartments. A target nucleic acid comprising genomic DNA is distributed into a first set of 96 vessels and the nucleic acids of each vessel are provided with a different first index using a Tn5-derived transposon. Thus a plurality of first-indexed template nucleic acids is obtained (e.g., Tn5-1, Tn5-2 ..., and Tn5-96). The plurality of first-indexed template nucleic acids are combined and then redistributed into a second set of 96 vessels and the nucleic acids of each vessel are provided with a different second index by amplification of the nucleic acids using primers comprising the second indexes. Thus a plurality of second-indexed template nucleic acids is obtained (e.g., PCR1, PCR2 ..., and PCR96). The plurality of second-indexed template nucleic acids can be combined and sequence information obtained. The use of 96 X 96 physical vessels is equivalent to 9216 virtual compartments.

<u>Methods to obtain extended haplotype information</u>

[0174] As described above, insertion of non-contiguous transposon sequences into the template nucleic acid can reduce positional information for a particular nucleic acid molecule, for example, after distribution of the template nucleic acids between several compartments. However, applicant has discovered methods to preserve such positional information for a particular nucleic acid molecule. Without being bound to any one theory, it has been observed that after transposition, the resulting two adjacent fragments of a particular nucleic acid molecule will tend to be distributed into the same vessel under conditions that maintain the transposase at the site of insertion of a transposon sequence. In other words, the transposase may hold the two resulting two adjacent fragments of a particular nucleic acid molecule together.

[0175] In some embodiments, a transposase can be removed from a template nucleic acid subsequent to distributing the template nucleic in several vessels. A transposase can be removed from the site of an insertion by various methods well known in the art, including the addition of a detergent, such as SDS, changing temperature, Proteinase digestion, chaperone capture and changing pH. DNA polymerases, with or without strand displacement properties including, but not limited to, phi29 DNA polymerase, Bst DNA polymerase, etc. can also be used to dislodge the transposase from the DNA.

[0176] FIG. 5 depicts an example scheme in which a target nucleic acid is distributed into a set of first vessels and indexed by insertion of transposomes, such as transposomes comprising non-contiguous transposon sequences. The first indexed template nucleic acids are pooled and distributed into a set of second vessels and indexed by PCR amplification. Sequence information can be obtained from the second indexed template nucleic acids.

[0177] Some methods of obtaining extended haplotype information from a target nucleic acid include (a) obtaining a template nucleic acid comprising a plurality of transposomes inserted into the target nucleic acid, wherein at least some of the inserted transposome each comprise a first transposon sequence, a second transposon sequence noncontiguous with the first transposon sequence, and a transposase associated with the first transposon sequence and the second transposon sequence; (b) compartmentalizing the template nucleic acid comprising the plurality of inserted transposomes into each vessel of a plurality of vessels; (c) removing the transposase from the template nucleic acid; and (d) obtaining sequence information from the template nucleic acid of each vessel, thereby obtaining haplotype information from the target nucleic acid. In some embodiments, compartmentalizing the template nucleic acid includes providing each vessel with an amount of template nucleic acid equivalent to greater than about a haploid equivalent of the target nucleic acid, an amount of template nucleic acid equivalent to about one haploid equivalent of the target nucleic acid, or an amount of template nucleic acid equivalent to less than about a haploid equivalent of the target nucleic acid.

[0178] An additional embodiment for maintaining contiguity of target nucleic acids for sequencing applications comprises utilizing one-sided (i.e., one transposon end) transpositional events in lieu of two-sided (i.e., two transposon ends) transpositional events as

disclosed herein. For example, transposases including, but not limited to Mu, MuE392Q mutant, Tn5 have been shown to display one-sided transposition of a transposon sequence into a target nucleic acid (Haapa et al., 1999, Nucl. Acids Res. 27(3): 2777-2784). The one-sided transpositional mechanism of these transposases can be utilized in methods described here to maintain the contiguity of a sample for sequencing, for example to haplotype or assemble a target nucleic acid.

[0179] In one example of one-sided transposition into a target DNA the transposome, a Tn5 dimer transposase is associated with only one transposon sequence end. In preferred embodiments, the transposon end could further comprise additional sequences such as index sequences, barcodes, and/or primer sequences and the like which could be used, for example, to identify a sample, amplify or extend the target nucleic acid and align fragment sequences. The transposome complex associates with the target nucleic acid, in that case dsDNA. At the site of transposome association, the transposase cleaves that strand of the target DNA and inserts the transposon and any other additional sequences at the point of cleavage. The transposase remains associated with the target DNA until it is removes, for example after partitioning of the sample as described herein the transposase can be removed by degradation (e.g., use of SDS or other methods as described here). The target nucleic acid, in this case dsDNA, does not fragment after removal of the transposase, as such the transposon and any additional sequences can be incorporated into the target DNA without fragmenting the DNA. Once the transposase is removed, target amplification by any means known in the art, in this example single or multiple primer amplification (due to incorporation of multiple different primer sequencing included in one or more transposons) either exponential or linear such as targeted PCR or whole genome amplification (for example by multiple strand displacement), can be performed to create libraries for sequencing. As described herein with respect to the two-sided transposon sequences, a variety of different combinations of index, barcode, restriction endonuclease site(s), and/or primer sequence could be included as part of the transposon sequence depending on the needs of the user. As such, one-sided transposome complexes could also be utilized to maintain contiguity of a target nucleic acid for methods disclosed herein for determining the haplotype of a target nucleic acid.

[0180] One sided transposomes can also be created from the two transposon/transposase complexes or the looped transposon/transposase complexes disclosed herein. For example, one of the transposon sequences of a two transposon complex or one end of the lopped transposon could be, for example, chemically modified or blocked so that transposition would not occur, or would minimally occur, at that end. For example, a dideoxynucleotide, a hapten such as a biotin could be incorporated at the end of one of the transposon ends which would inhibit transposition at that end, thereby allowing for only one transposon, or one end of a looped transposon, to be inserted into the target nucleic acid.

[0181] In one embodiment, a method of obtaining sequence information from a target nucleic acid comprises obtaining a template nucleic acid comprising a plurality of transposons inserted into said target nucleic acid such that the contiguity of the template is retained, compartmentalizing the nucleic acids comprising the plurality of inserted transposons into a

plurality of vessels, generating compartment-specific indexed libraries from the transposed nucleic acid targets and obtaining sequence information from the template nucleic acids in each vessel of the plurality of vessels.

Certain methods for preparing target nucleic acids for haplotyping

[0182] Some embodiments of the methods and compositions provided herein include preparing target nucleic acids for haplotyping using the methods provided herein. Using a pre-amplification method, the number of unique reads is increased by generating multiple identical copies of the same nucleic acid fragment as a contiguous product. In some such embodiments, a library is amplified by methods such as rolling circle amplification (RCA). In some embodiments, circular libraries of a target nucleic acid are prepared and the library amplified by RCA. Such methods generate extended long nucleic acids.

[0183] An example scheme is shown in **FIG. 6. FIG. 6** depicts a method including preparing target nucleic acids for haplotyping by generating a library comprising circular molecules by mate-pair and selection of specific or range of sizes from 1-10kb or 10-20kb, or 20kb-50kb, 50-200kb nucleic acids; amplifying the library by RCA to generate extended lone nucleic acids; inserting indexes into the amplified library with transposons; compartmentalizing the inserted library; removing transposase with SDS; further indexing the library; and obtaining sequence information from the library.

[0184] Another example scheme is shown in **FIG. 7. FIG. 7** depicts a method including preparing target nucleic acids for haplotyping by generating a library comprising circular molecules by hairpin transposition, gap fill, and selection of a specific size or range of sizes from 1-10kb or 10-20kb, or 20kb-50kb, 50-200kb5nucleic acids; amplifying the library by RCA to generate extended lone nucleic acids; inserting indexes into the amplified library with transposons; compartmentalizing the inserted library; removing transposase with SDS; further indexing the library; and obtaining sequence information from the library.

Methods to generate mate-paired libraries

[0185] Methods for generating mate-pair libraries include: fragmenting genomic DNA into large fragments typically greater than (though not limited to) 1000 bp; circularizing individual fragments by a method that tags the ligated junction; fragmenting the DNA further; enriching the tagged junction sequences and ligating adaptors to the enriched junction sequences so that they may be sequenced yielding information about the pair of sequences at the ends of the original long fragment of DNA. These processes involve at least 2 steps where DNA is fragmented, either physically or enzymatically. In at least one or more distinct steps, adaptors are ligated to the ends of fragments. Mate-pair preps typically take 2-3 days to perform and comprise multiple steps of DNA manipulations. The diversity of the resulting library correlates

directly with the number of steps required to make the library.

[0186] The method provided herein simplifies the number of steps in the library generation protocol by employing a transposase mediated reaction that simultaneously fragments and adds adaptor sequences to the ends of the fragments. At least one or both of the fragmentation steps (initial fragmentation of genomic DNA and fragmentation of circularized fragments) may be performed with a transposome, thus replacing the need for separate fragmentation and adaptor ligation step. Obviating the polishing, preparation of, and ligation to fragment ends reduces the number of process steps and thus increases the yield of usable data in the prep as well as making the procedure more robust. In one embodiment, the protocol can be performed without resorting to methods that purify a selection of sizes based on electrophoresis. This method produces a broader range of fragment sizes than can be achieved with gel electrophoretic methods but nonetheless produces usable data. The advantage is that a labour intensive step is avoided.

[0187] FIG. 8 provides an example scheme where just the initial fragmentation is replaced with a transposome tagmentation step. The circularized DNA is fragmented by either physical methods or chemical/enzymatic methods and the fragments turned into a library via application of standard sample prep protocols (*e.g.* TRUSEQ). **FIG. 9** illustrates an example scheme where both the initial fragmentation and the circularized DNA fragmentation are performed with a transposome tagmentation step. The adaptor sequences for the transposome (including the ME sequences) may or may not be different for the adaptor used for the initial tag mentation and the subsequent circle tagmentation.

[0188] Amplifying template nucleic acid by generating multiple copies of each molecule before transposition or introduction of molecular indexes creates redundancy which can be useful for getting higher SNP coverage in each haplotype block, and also for de novo genome assembly, similar to a shotgun approach. Template nucleic acid can be converted to a defined-size library by low-frequency transposition, physical shearing, or enzymatic digestion, and then amplified for a finite number of cycles by either PCR or a whole genome amplification scheme (for example using phi29). The amplified library which already contains the built-in redundancy can be used as the input material for the haplotyping workflow. This way, every region of the genome is represented multiple times by multiple copies generated upfront with each copy contributing a partial coverage of that region; however, the consensus coverage will be closer to complete.

EXAMPLES

Example 1-Reducing error rates

[0189] A library of template nucleic acids was prepared with each fragment comprising a

different barcode. Each fragment was amplified and sequence information was obtained from at least one amplified product from each fragment. A consensus sequence was determined from the sequence information from the amplified products from each fragment. In particular, a NEXTERA-prepped sequencing library was sequenced for 500 cycles on a MISEQ instrument. The library consisted of a distribution of sizes, with maximum read lengths extending to -300 nt. The error rates at cycle 250 were approximately 15%. If a template was represented just three times, the error rate dropped to ~1% at cycle 250. **FIG. 10** illustrates a model of error rates with number of amplified products sequenced (coverage of each barcode). Error rates decrease as the number of amplified products sequenced from a fragment increase.

Example 2-Coupling transposon sequences

[0190] This example illustrates methods to couple two transposon sequences together in various orientations including a 5'-5' orientation, and a 3'-3' orientation. In an exemplary method, aldehyde oxyamine is used to form linked oligos via oxime ether formation. An aldehyde modified oligo (either on the 5' or the 3' end) is combined with an oligo modified with an oxyamine on the 5' end in reaction buffer and allowed to incubate for 2 hours. Final product can be isolated via PAGE purification for example.

[0191] In another exemplary method, bisoxyamine coupling was performed. Aldehyde modified oligos were dimerized with a bis-oxyamine (e.g., dioxyamino butane) linker using locally high concentrations to force bi-substitution. 100-mer oligos were synthesized with an aldehyde on the 5' end and purified. The bisoxyamine oligo was synthesized. A 1 mM solution of the bisoxyamine oligo was made in a low pH reaction buffer containing a catalyst (5 M urea, 100 mM aniline, 10 mM citrate, 150 mM NaCl, pH 5.6) and added to a 665 μM solution of aldehyde oligo in water. The entire volume of the solution was diluted 1:1 with reaction buffer, and allowed to incubate at room temperature for 2 hours. A titration of various aldehyde:bisoxyamine ratios showed dimerization at high bisoxyamine ratios. The most successful conditions were replicated with 3' aldehyde oligos. **FIG. 11** shows results of 5'-5' bisoxyamine coupling reactions in which looped precursor transposons were observed in the indicated dimer band. Similar results were observed for 3'-3' bisoxyamine coupling reactions.

Example 3-Monitoring transposome stability

[0192] Transposomes were prepared using long and short transposon sequences loaded onto transposase. The transposome products included: A (2 short sequences); B (long and short sequences); and C (2 long sequences). The relative amounts of each species of transposome were measured under various conditions, such as temperature, buffers, ratios of transposon sequences to transposase. Generally, high NaCl or KCl salt increased exchange of transposon sequences between transposomes. Glutamate and acetate buffers eliminated or reduced exchange, with preferred concentrations between 100-600 mM. Optimum storage conditions

were determined.

Example 4-Maintaining template contiguity

[0193] This example illustrates a method for maintaining contiguity information of a template nucleic acid prepared using transposomes comprising non-contiguous transposon sequences in which Tn5 transposase stays bound to the template DNA post-transposition. Target nucleic acid was contacted with transposomes comprising Tn5 transposase, and non-contiguous transposon sequences. **FIG. 12** shows that samples further treated with SDS appeared as a smear of various fragments of template nucleic acid; samples not treated with SDS showed retention of putative high molecular weight template nucleic acid. Thus, even though a nucleic acid may be fragmented, adjacent sequences may still be associated with one another by the transposase (as demonstrated by the Tn5 bound DNA left in the wells).

[0194] In still another exemplary method, a library of template nucleic acids was prepared using transposomes comprising non-contiguous transposon sequences with target nucleic acid comprising human Chromosome 22. **FIG. 13** summarizes that haplotype blocks up to 100 kb were observed for samples in which transposase was removed by SDS post-dilution. Thus, by practicing methods as described herein target nucleic acids can maintain target integrity when transposed, be diluted, and be transformed into sequencing libraries.

Example 5-Maintaining template contiguity

[0195] Target nucleic acids were tagmented with transposomes comprising non-contiguous transposon sequences (NEXTERA), diluted to the desired concentration, and then treated with SDS to remove the transposase enzyme before PCR. As a control, the same amount of input DNA was tagmented, treated with SDS first and then diluted to the desired concentration. SDS treatment before dilution removes proximity information since the transposase enzyme dissociates from the target DNA with SDS, thereby fragmenting the target DNA. Two tagmentation reactions were set-up on 50ng of a Coriell gDNA, 1 reaction was stopped with 0.1% SDS and diluted to 6pg. Next, the other reaction was first diluted to 6pg and then stopped with 0.1% SDS. The entire reactions were used to set up a 30-cycle PCR reaction and sequenced on a Gene Analyzer platform (Illumina), according to the manufacturer's instructions. The reads were mapped to a human reference genome and the distance distribution was calculated and plotted.

[0196] As shown in **FIG. 14**, in the SDS-PostDilution sample, the median distance shifted to smaller sizes and a large sub-population of proximally located reads becomes apparent. If there was any haplotyping, pile up of proximal reads was expected. The bi-modal distribution of post-dilution sample demonstrated that there is an enrichment of proximal reads.

[0197] The distance distribution was a measure of sample size (i.e. the more unique reads, the shorter the distance). The distance histogram for SDS-PreDilution and SDS-PostDilution samples are shown **in FIG. 14**. To correct for the difference in the number of unique reads, the Pre-Dilution was down-sampled to give the same number of unique mapped reads (664,741). A significant enrichment was observed for reads that are immediate neighbors (i.e. junctions). This was measured by looking at the "distance to next alignment' distribution and finding the reads which their distance to their next alignment is the sequence read length minus 9 (which correspond to 9bp duplication caused by Tn5 at the insertion site). Such reads make up 10% of the data (**FIG. 15**) and with employment of single primer system for amplification of NEXTERA libraries, can be doubled. Also implementation of a more conservative sample prep which allows less sample loss allows recovery of more junction data. The haplotype resolving power diminished when input DNA was increased. On the other hand, reducing the DNA input required more amplification, and therefore more PCR cycles, generating many PCR duplicates. Using individually barcoded Tn5 complexes allowed the tagmentation and subsequent dilutions to be carried out in separate compartments. Low levels of input from individually barcoded and tagmented material were combined to elevate the PCR input DNA amounts to the level that allowed more specific amplification with less waste of sequencing capacity on redundant reads. Accordingly, using sufficient barcoded complexes allowed phasing of the majority of the human genome. In order to increase the haplotype resolving power, barcoding was implemented at both the complex level and PCR primer level. Such combinatorial indexing scheme allows the use of very low input DNA from each individually barcoded complex into PCR reaction, which would allow powerful haplotype resolution. Using only 40 indexing oligos (8+12=20 for NEXTERA complexes which generates 8* 12=96 individual complexes and 8+12=20 for PCR primers which would allow 8*12=96 additional indexes), 96*96=9216 virtual compartments were generated for the abovementioned haplotyping workflow. Using a modified sequencing recipe, all the data was sequenced on a HiSeq-2000. All 9216 possible barcode combinations were observed in the sequencing results.

Example 6-Obtaining haplotype information with Mu

[0198] Transposomes comprising Mu were used to obtain haplotype information. 1 ng of genomic DNA was targeted with Mu-TSM in a 50 µl reaction volume with 1 X TA buffer and 1, 2, 4, or 8 µl of 25 µM Mu-TSM complexes. Reactions were incubated at 37 °C for 2 hours. Samples were diluted to 1 pg/µl. For Mu inactivation, 10µl of each sample containing either 1 pg or 5 pg total genomic DNA were prepared. SDS was added to final concentration of 0.05%. Samples were incubated at 55 °C for 20 minutes. The whole sample was used to set-up a 50 µl PCR reaction using NPM. PCR was done for 30 cycles. PCR samples were cleaned up with 0.6 X SPRI and resuspended in 20 µl of re-suspension buffer. Sequencing information was obtained. FIG. 16 shows the observed Pile-up proximal reads observed at sub-haploid content using transposomes comprising Mu.

[0199] 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.

REFERENCES CITED IN THE DESCRIPTION

This list of references cited by the applicant is for the reader's convenience only. It does not form part of the European patent document. Even though great care has been taken in compiling the references, errors or omissions cannot be excluded and the EPO disclaims all liability in this regard.

Patent documents cited in the description

- US20120208705A [0073] [0079] [0093] [0103] [0106] [0107] [0108] [0121] [0131]
- US20120208724A [0073] [0079] [0093] [0103] [0106] [0107] [0108] [0121] [0131]
- WO2012061832A [0073] [0079] [0093] [0103] [0106] [0107] [0108] [0121] [0131]
- US5644048A [0081]
- US5386023A [0082]
- US5637684A [0082]
- US5602240A [0082]
- US5216141A [0082]
- US4469863A [0082]
- US5235033A [0082]
- US5034506A [0082] [0085]
- US5432272A [0085]
- WO9523875A [0091]
- US20100120098A [0094] [0171]
- US7741463B [0105]
- WO2012108864A [0114]
- US7582420B [0136]
- US7955794B [0136]
- US8003354B [0136]
- WO2010062913A [0141]
- US20100323348A [0143]
- US7574305B [0143]
- US6654696B [0143]
- US6210891B [0147]
- US6258568B [0147]
- US6274320A [0147]
- US742767A [0148]
- US74141163B [0148]
- US7057026B [0148] [0149]

- WO9106678A [0148]
- WO07123744A [0148]
- US20070166705A [0149]
- US20060188901A [0149]
- US20060240439A [0149]
- US20060281109A [0149]
- WO05065814A [0149]
- US20050100900A [0149]
- WO06064199A [0149]
- WO07010251A [0149]
- US6969488B [0150]
- US6172218B [0150]
- US6306597B [0150]
- US7001792B [0151]
- US7329492B [0152]
- US7211414B [0152]
- US7315019B [0152]
- US7405281B [0152]
- US20080108082A [0152]
- US7181122B [0152]
- US7302146B [0152]
- US7313308B [0152]

Non-patent literature cited in the description

- **BEAUCAGE et al.**Tetrahedron, 1993, vol. 49, 1925- [0081]
- **LETSINGER, J**Org. Chem., 1970, vol. 35, 3800- [0081]
- **SPRINZL et al.**Eur. J. Biochem., 1977, vol. 81, 579- [0081]
- **LETSINGER et al.**Nucl. Acids Res., 1986, vol. 14, 3487- [0081]
- **SAWAI et al.**Chem. Lett., vol. 805, 1984- [0081]
- **LETSINGER et al.**J. Am. Chem. Soc., 1988, vol. 110, 4470- [0081] [0082]
- **PAUWELS et al.**Chemica Scripta, 1986, vol. 26, 141- [0081]
- **MAG et al.**Nucleic Acids Res., 1991, vol. 19, 1437- [0081]
- **BRIU et al.**J. Am. Chem. Soc., 1989, vol. 111, 2321- [0081]
- **ECKSTEIN**Oligonucleotides and Analogues: A Practical ApproachOxford University Press [0081]
- **EGHOLM, J.**Am. Chem. Soc., 1992, vol. 114, 1895- [0081]
- **MEIER et al.**Chem. Int. Ed. Engl., 1992, vol. 31, 1008- [0081]
- **NIELSEN**Nature, 1993, vol. 365, 566- [0081]
- **CARLSSON et al.**Nature, 1996, vol. 380, 207- [0081]

- **DENPCY et al.**Proc. Natl. Acad. Sci. USA, 1995, vol. 92, 6097- [0082]
- **KIEDROWSHI et al.**Angew. Chem. Intl. Ed. English, 1991, vol. 30, 423- [0082]
- **LETSINGER et al.**Nucleosides & Nucleotides, 1994, vol. 13, 1597- [0082]
- Carbohydrate Modifications in Antisense ResearchASC Symposium Series 580 [0082] [0082]
- **MESMAEKER et al.**Bioorganic & Medicinal Chem. Lett., 1994, vol. 4, 395- [0082]
- **JEFFS et al.**J. Biomolecular NMR, 1994, vol. 34, 17- [0082]
- Tetrahedron Lett., 1996, vol. 37, 743- [0082]
- **JENKINS et al.**Chem. Soc. Rev., 1995, 169-176 [0082]
- **LOAKES et al.**Nucleic Acid Res., 1994, vol. 22, 4039- [0084]
- **AERSCHOT et al.**Nucleic Acid Res., 1995, vol. 23, 4363- [0084]
- **NICHOLS et al.**Nature, 1994, vol. 369, 492- [0084]
- **BERGSTROM et al.**Nucleic Acid Res., 1997, vol. 25, 1935- [0084]
- **LOAKES et al.**Nucleic Acid Res., 1995, vol. 23, 2361- [0084]
- **LOAKES et al.**J. Mol. Biol., 1997, vol. 270, 426- [0084]
- **FOTIN et al.**Nucleic Acid Res., 1998, vol. 26, 1515- [0084]
- **SCHEIT**Nucleotide AnalogsJohn Wiley19800000 [0085]
- **ENGLISCH**Angew. Chem. Int. Ed. Engl., 1991, vol. 30, 613-29 [0085]
- **AGARWAL**Protocols for Polynucleotides and AnalogsHumana Press19940000 [0085]
- **S. VERMAF. ECKSTEIN**Ann. Rev. Biochem., 1998, vol. 67, 99-134 [0085]
- **T**he Glen Report, 2003, vol. 16, 25- [0085]
- **KOSHKIN et al.T**etrahedron, 1998, vol. 54, 3607-30 [0085]
- **STERCHAK, E. P. et al.**Organic Chem., 1987, vol. 52, 4202- [0085]
- **NIELSEN et al.**Science, 1991, vol. 254, 1497-1500 [0085]
- **EGHOLM et al.**J. Am. Chem. Soc., 1992, vol. 114, 1895-1897 [0085]
- **DEMIDOV et al.**Proc. Natl. Acad. Sci., 2002, vol. 99, 5953-58 [0085]
- Peptide Nucleic Acids: Protocols and ApplicationsHorizon Bioscience20040000 [0085]
- **GORYSHINREZNIKOFF**J. Biol. Chem., 1998, vol. 273, 7367- [0090]
- **MIZUUCHI, K.**Cell, 1983, vol. 35, 785- [0090]
- **SAVILAHTI, H et al.**EMBO J., 1995, vol. 14, 4893- [0090]
- **COLEGIO et al.**J. Bacteriol., 2001, vol. 183, 2384-8 [0091]
- **KIRBY C et al.**Mol. Microbiol., 2002, vol. 43, 173-86 [0091]
- **DEVINEBOEKE**Nucleic Acids Res., 1994, vol. 22, 3765-72 [0091]
- **CRAIG, N L**Science, 1996, vol. 271, 1512- [0091]
- **CRAIG, N L**Curr Top Microbiol Immunol., 1996, vol. 204, 27-48 [0091]
- **KLECKNER N et al.**Curr Top Microbiol Immunol., 1996, vol. 204, 49-82 [0091]
- **LAMPE D J et al.**EMBO J., 1996, vol. 15, 5470-9 [0091]
- **PLASTERK R H**Curr. **T**opics Microbiol. Immunol., 1996, vol. 204, 125-43 [0091]
- **GLOOR, G B**Methods Mol. Biol., 2004, vol. 260, 97-114 [0091]
- **ICHIKAWAOHTSUBO**J Biol. Chem., 1990, vol. 265, 18829-32 [0091]
- **OHTSUBOSEKINE**Curr. **T**op. Microbiol. Immunol., 1996, vol. 204, 1-26 [0091]
- **BROWN et al.**Proc Natl Acad Sci USA, 1989, vol. 86, 2525-9 [0091]
- **BOEKECORCES**Annu Rev Microbiol., 1989, vol. 43, 403-34 [0091]
- **ZHANG et al.**PLoS Genet., 2009, vol. 5, e1000689- [0091]

- **WILSON C. et al.**J. Microbiol. Methods, 2007, vol. 71, 332-5 [0091]
- **RONAGHI et al.**Analytical Biochemistry, 1996, vol. 242, 184-9 [0147]
- **RONAGHI, M.**Genome Res., 2001, vol. 11, 13-11 [0147]
- **RONAGHI et al.**Science, 1998, vol. 281, 5375363- [0147]
- **DEAMER, D.W.AKESON, M.**Nanopores and nucleic acids: prospects for ultrarapid sequencingTrends Biotechnol., 2000, vol. 18, 147-151 [0151]
- **DEAMER, D.D. BRANTON**Characterization of nucleic acids by nanopore analysisAcc. Chem. Res., 2002, vol. 35, 817-825 [0151]
- **LI et al.**DNA molecules and configurations in a solid-state nanopore microscopeNat. Mater., 2003, vol. 2, 611-615 [0151]
- **SONIMELLER**A. Progress toward ultrafast DNA sequencing using solid-state nanoporesClin. Chem., 2007, vol. 53, 1996-2001 [0151]
- **HEALY, K.**Nanopore-based single-molecule DNA analysisNanomed., 2007, vol. 2, 459-481 [0151]
- **COCKROFT et al.**A single-molecule nanopore device detects DNA polymerase activity with single-nucleotide resolutionJ. Am. Chem. Soc., 2008, vol. 130, 818-820 [0151]
- **LEVENE, M.J. et al.**Zero-mode waveguides for single-molecule analysis at high concentrationsScience, 2003, vol. 299, 682-686 [0152]
- **LUNDQUIST, P.M. et al.**Parallel confocal detection of single molecules in real timeOpt. Lett., 2008, vol. 33, 1026-1028 [0152]
- **KORLACH et al.**Selective aluminum passivation for targeted immobilization of single DNA polymerase molecules in zero-mode waveguide nanostructuresProc. Natl. Acad. Sci. USA, 2008, vol. 105, 1176-1181 [0152]
- **KORLACH J. et al.**Long, processive enzymatic DNA synthesis using 100% dye-labeled terminal phosphate-linked nucleotidesNucleosides, Nucleotides and Nucleic Acids, 2008, vol. 27, 1072-1083 [0153]
- **HARRIS T.D. et al.**Single Molecule DNA Sequencing of a viral GenomeScience, 2008, vol. 320, 106-109 [0154]
- **KINDE I. et al.**PNAS, 2011, vol. 108, 9530-9535 [0162]
- **SCHMITT M.W. et al.**PNAS, 2012, vol. 109, 14508-13 [0165]
- **HAAPA et al.**Nucl. Acids Res., 1999, vol. 27, 32777-2784 [0178]

PATENTKRAV

1. Fremgangsmåde til fremstilling af et bibliotek af template-nukleinsyrer til opnåelse af sekvensinformationer fra en målnukleinsyre, hvilken fremgangsmåde omfatter:

(a) kompartmentalisering af målnukleinsyren i en flerhed af første kar ved forsyning af hvert første kar med en mængde målnukleinsyre på mere end ca. én eller flere haploidækvivalenter af målnukleinsyren;

(b) tilvejebringelse af et første indeks til målnukleinsyren i hvert første kar, hvor det første indeks, der tilvejebringes til målnukleinsyren i hvert første kar, er forskelligt, hvor trinnet omfatter, at målnukleinsyren bringes i kontakt med en flerhed af transposomer, hvor hvert transposom omfatter en første transposonsekvens, der omfatter et første indeks, en anden transposonsekvens, som er ikke-sammenhængende med den første transposonsekvens, og en transposase, som er forbundet med den første transposonsekvens og den anden transposonsekvens under sådanne betingelser, at mindst nogle af transposonsekvenserne indsættes i målnukleinsyren, hvorved der opnås en flerhed af første indekserede template-nukleinsyrer;

(c) kombination af de første indekserede template-nukleinsyrer;

(d) kompartmentalisering af de første indekserede template-nukleinsyrer i en flerhed af sekundære kar; og

(e) tilvejebringelse af et andet indeks til de første indekserede template-nukleinsyrer i hvert sekundært kar, hvor det andet indeks, der tilvejebringes til de første indekserede template-nukleinsyrer i hvert sekundært kar, er forskelligt, hvorved der opnås en flerhed af sekundære indekserede template-nukleinsyrer.

2. Fremgangsmåde ifølge krav 1, hvor trin (d) omfatter fjernelse af transposasen fra de kompartmentaliserede første indekserede template-nukleinsyrer, hvor fjernelse af transposasen omfatter en fremgangsmåde, der er valgt fra gruppen bestående af tilsætning af et detergent, ændring af temperatur, ændring af pH-værdi, tilsætning af en protease, tilsætning af en chaperone og tilsætning af en polymerase.

3. Fremgangsmåde ifølge et hvilket som helst af kravene 1-2, hvor den første transposonsekvens omfatter et første primersted, og den anden transposonsekvens omfatter et andet primersted.

4. Fremgangsmåde ifølge krav 3, hvor det første primersted endvidere omfatter en første stregkode, og det andet primersted endvidere omfatter en anden stregkode, hvor den første stregkode og den anden stregkode er forskellige.

5. Fremgangsmåde ifølge et hvilket som helst af kravene 1-4, hvor trin (d) omfatter forsyning af hvert første kar med en mængde af den første indekserede template-nukleinsyre på mere end ca. én eller flere haploidækvivalenter af målnukleinsyren.

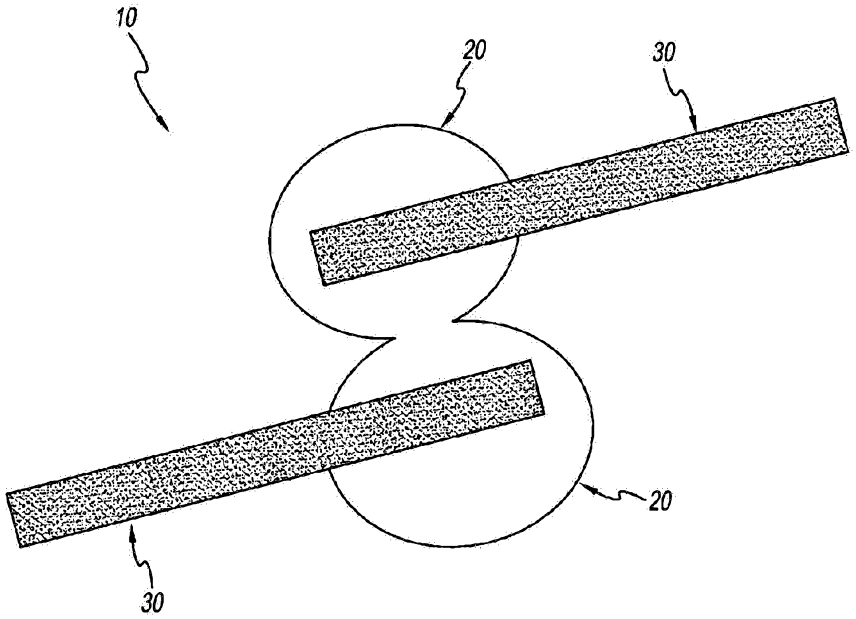
6. Fremgangsmåde ifølge et hvilket som helst af kravene 1-5, hvor trin (e) omfatter amplificering af de første indekserede template-nukleinsyrer med mindst én primer, der omfatter det andet indeks.

7. Fremgangsmåde ifølge et hvilket som helst af kravene 1-5, hvor trin (e) omfatter ligering af de første indekserede template-nukleinsyrer med mindst én primer, der omfatter det andet indeks.

8. Fremgangsmåde ifølge et hvilket som helst af kravene 1-7, hvor målnukleinsyren opnås ved at berige en flerhed af nukleinsyrer for en sekvens af interesse enten inden eller efter transposition.

9. Fremgangsmåde ifølge krav 1, der endvidere omfatter opnåelse af sekvensinformationer fra template-nukleinsyren, hvor sekvensinformationerne valgfrit omfatter sekvensinformationer af haplotypen.

DRAWINGS



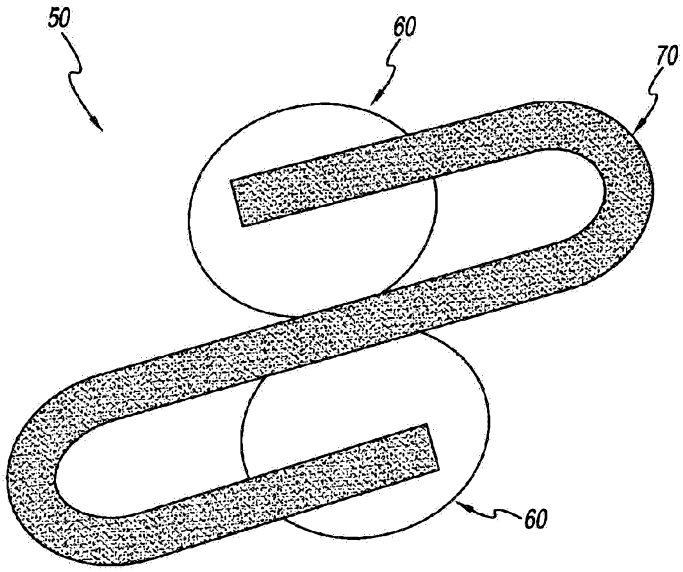


FIG. 1

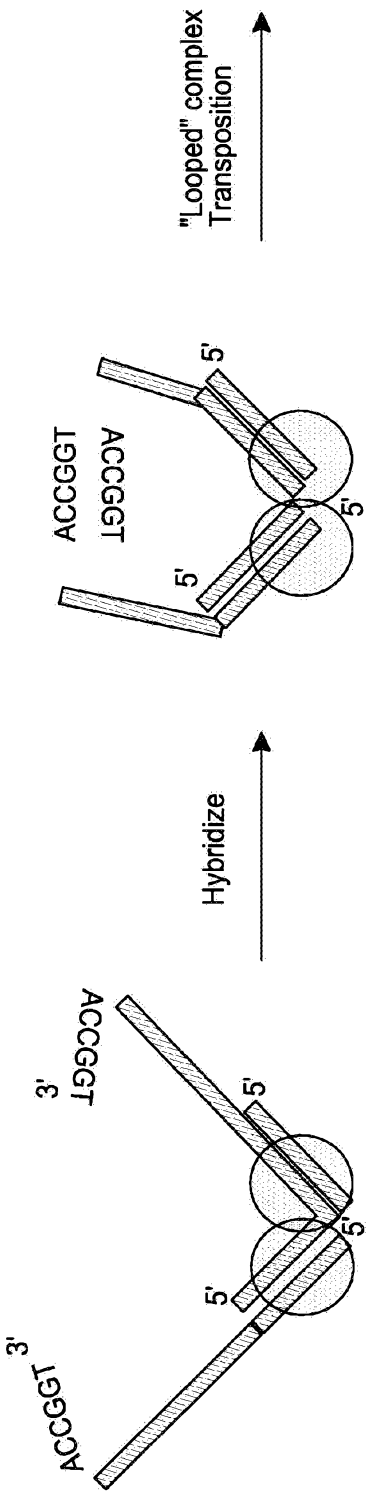

3'
ACCGGT
5'
5'
3'
ACCGGT
Hybridize
5'
5'
ACCGGT
ACCGGT
5'
5'
"Looped" complex
Transposition

FIG. 2

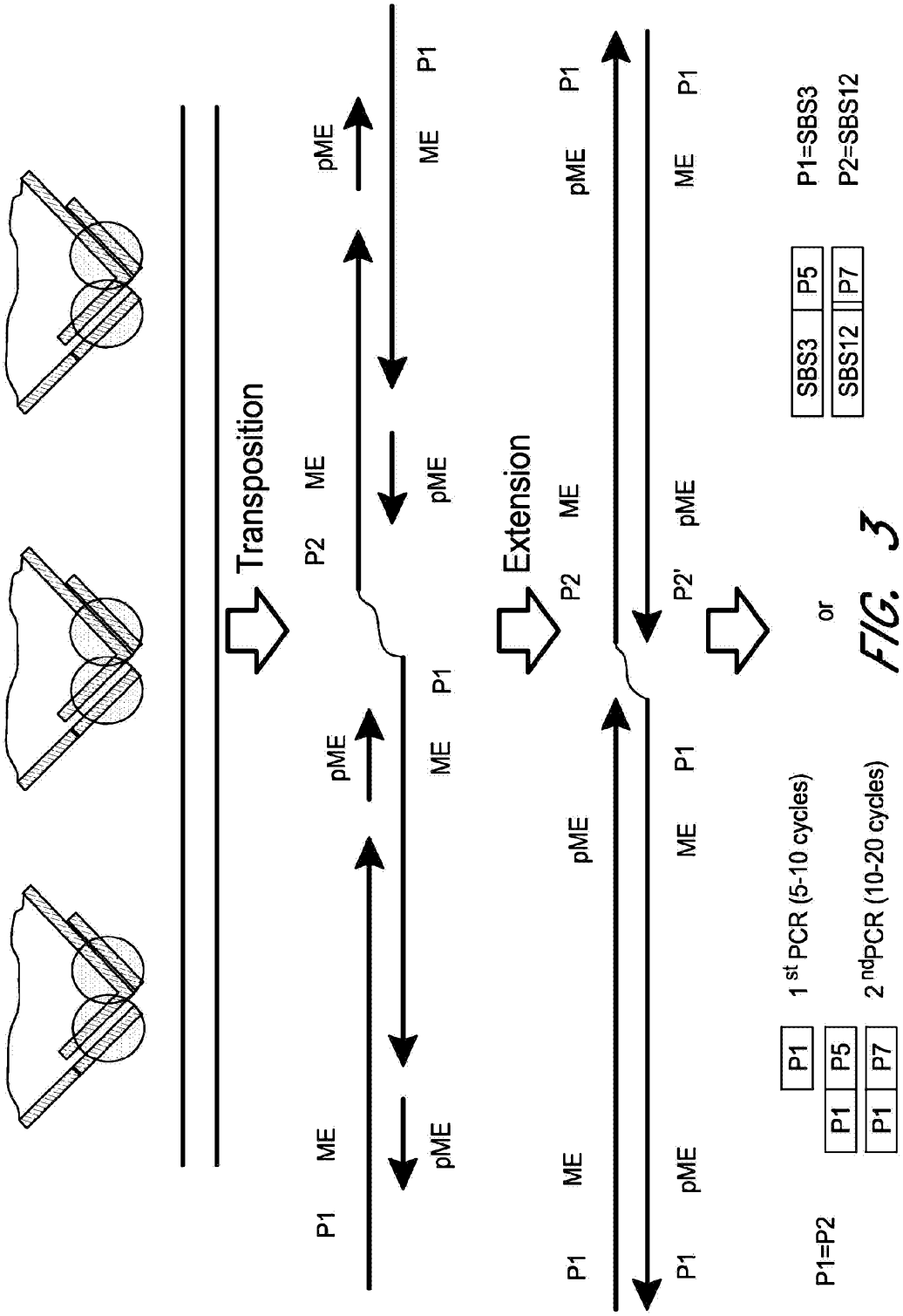

Transposition
Extension
P1
ME
pME
P2
P2'
P1=P2
P1
P5
P7
1 st PCR (5-10 cycles)
2 ndPCR (10-20 cycles)
or
SBS3
SBS12
P1=SBS3
P2=SBS12

FIG. 3

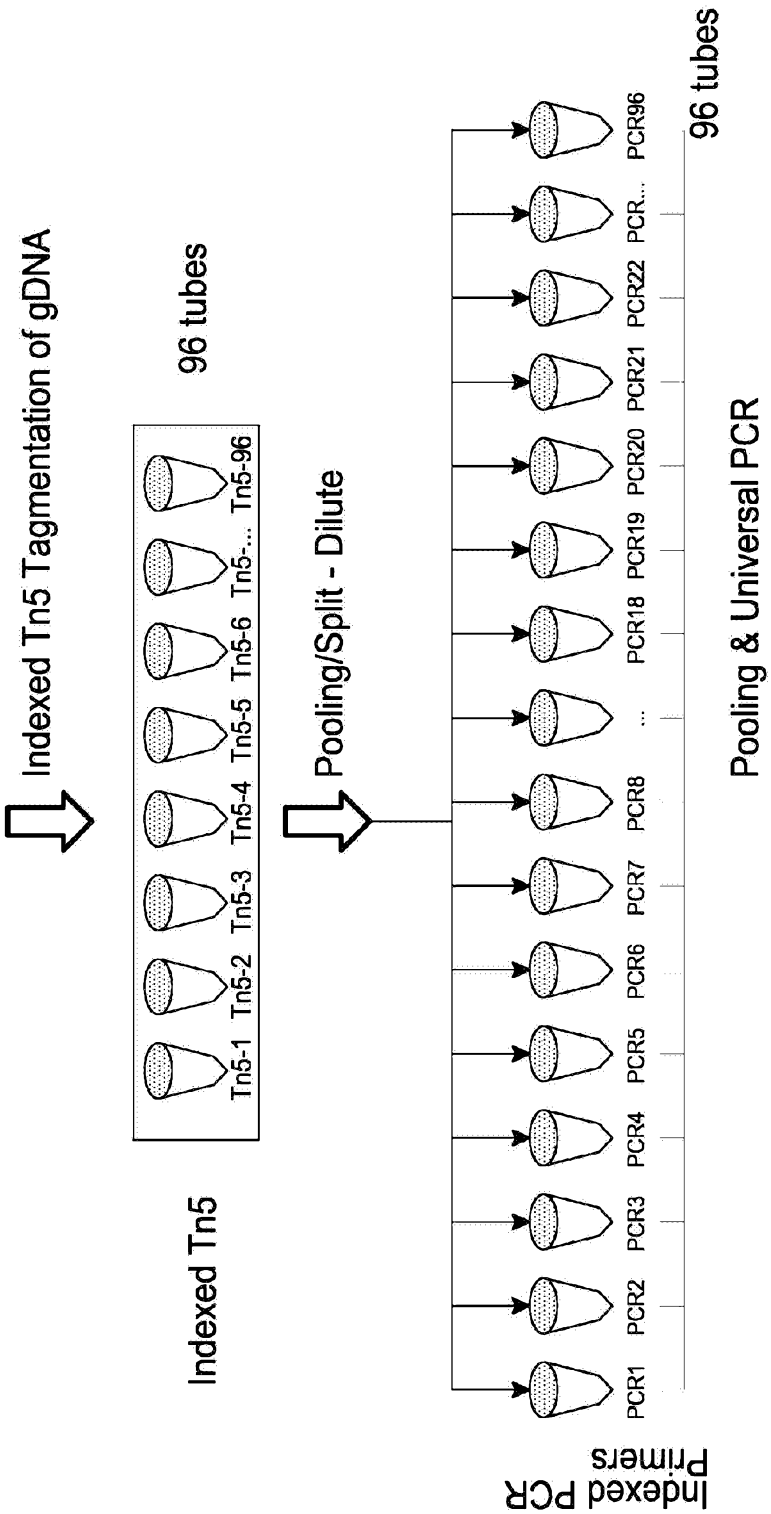

Indexed Tn5 Tagmentation of gDNA
96 tubes
Indexed Tn5
Tn5-1
Tn5-2
Tn5-3
Tn5-4
Tn5-5
Tn5-6
Tn5-...
Tn5-96
Pooling/Split - Dilute
Indexed PCR Primers
PCR1
PCR2
PCR3
PCR4
PCR5
PCR6
PCR7
PCR8
...
PCR18
PCR19
PCR20
PCR21
PCR22
PCR...
PCR96
96 tubes
Pooling & Universal PCR

FIG. 4

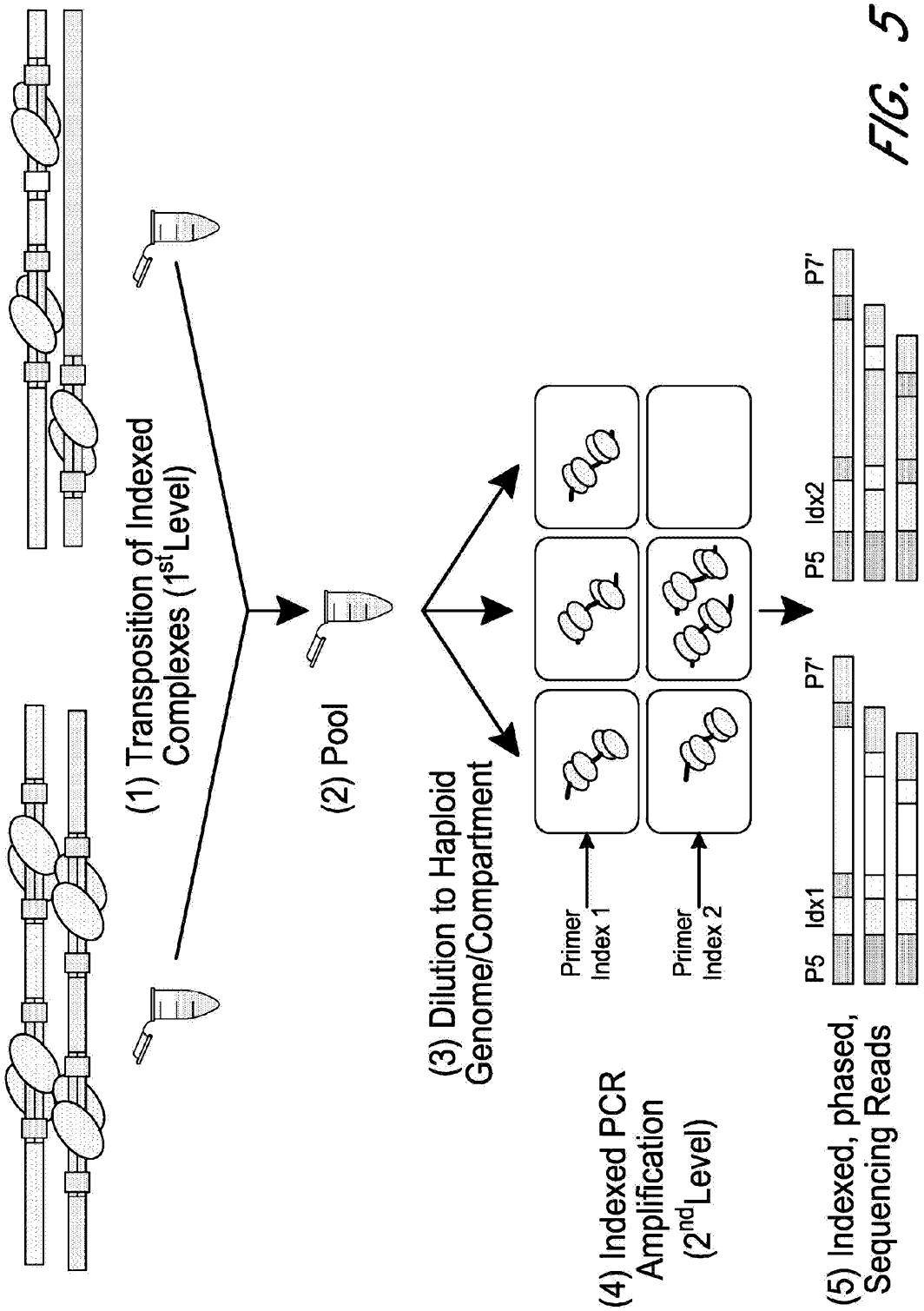

(1) Transposition of Indexed Complexes (1st Level)
(2) Pool
(3) Dilution to Haploid Genome/Compartment
Primer Index 1
Primer Index 2
(4) Indexed PCR Amplification (2nd Level)
P5
Idx1
P7'
P5
Idx2
P7'
(5) Indexed, phased, Sequencing Reads

FIG. 5

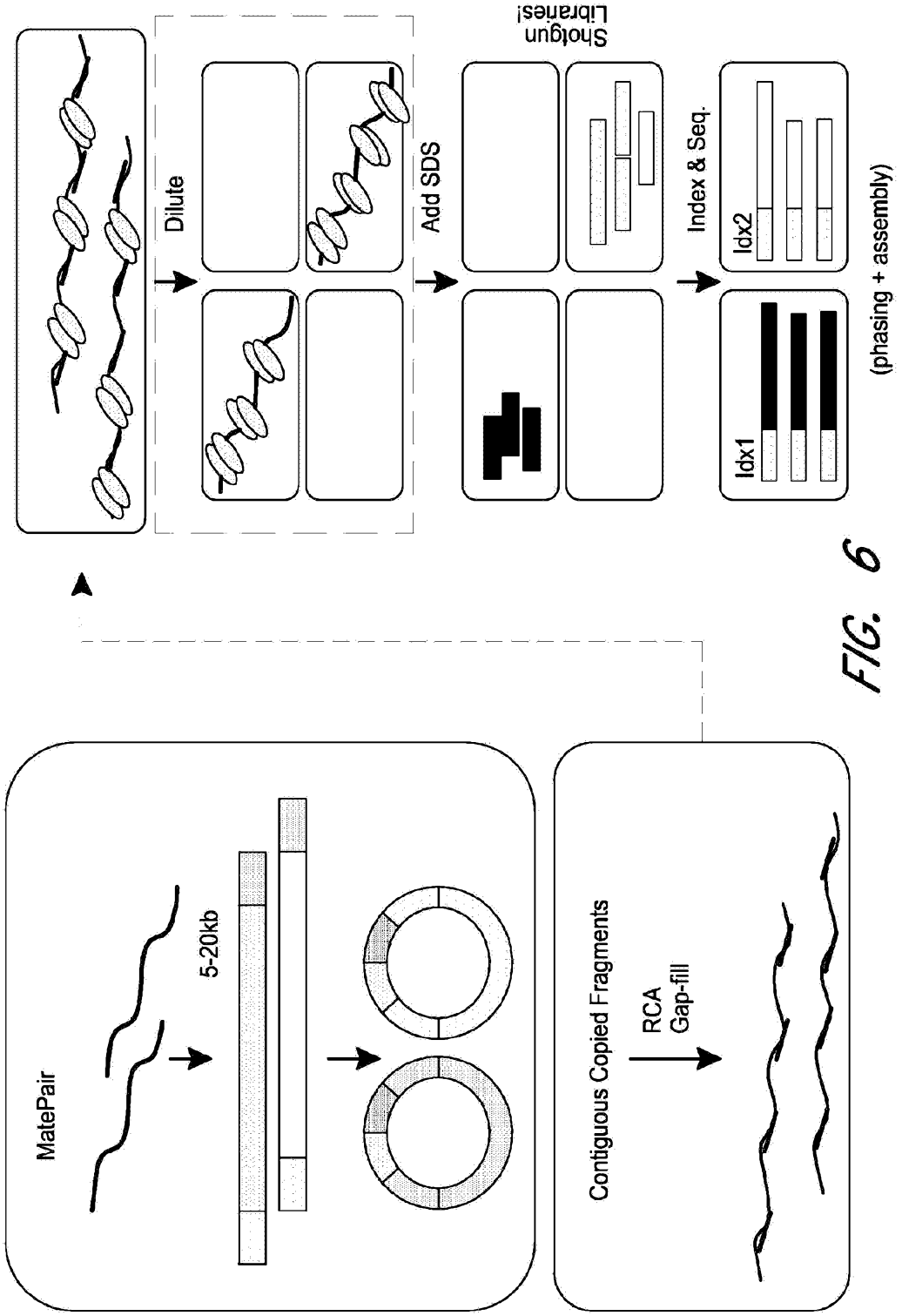

MatePair
5-20kb
Contiguous Copied Fragments
RCA
Gap-fill
Dilute
Add SDS
Shotgun Libraries!
Index & Seq.
Idx1
Idx2
(phasing + assembly)

FIG. 6

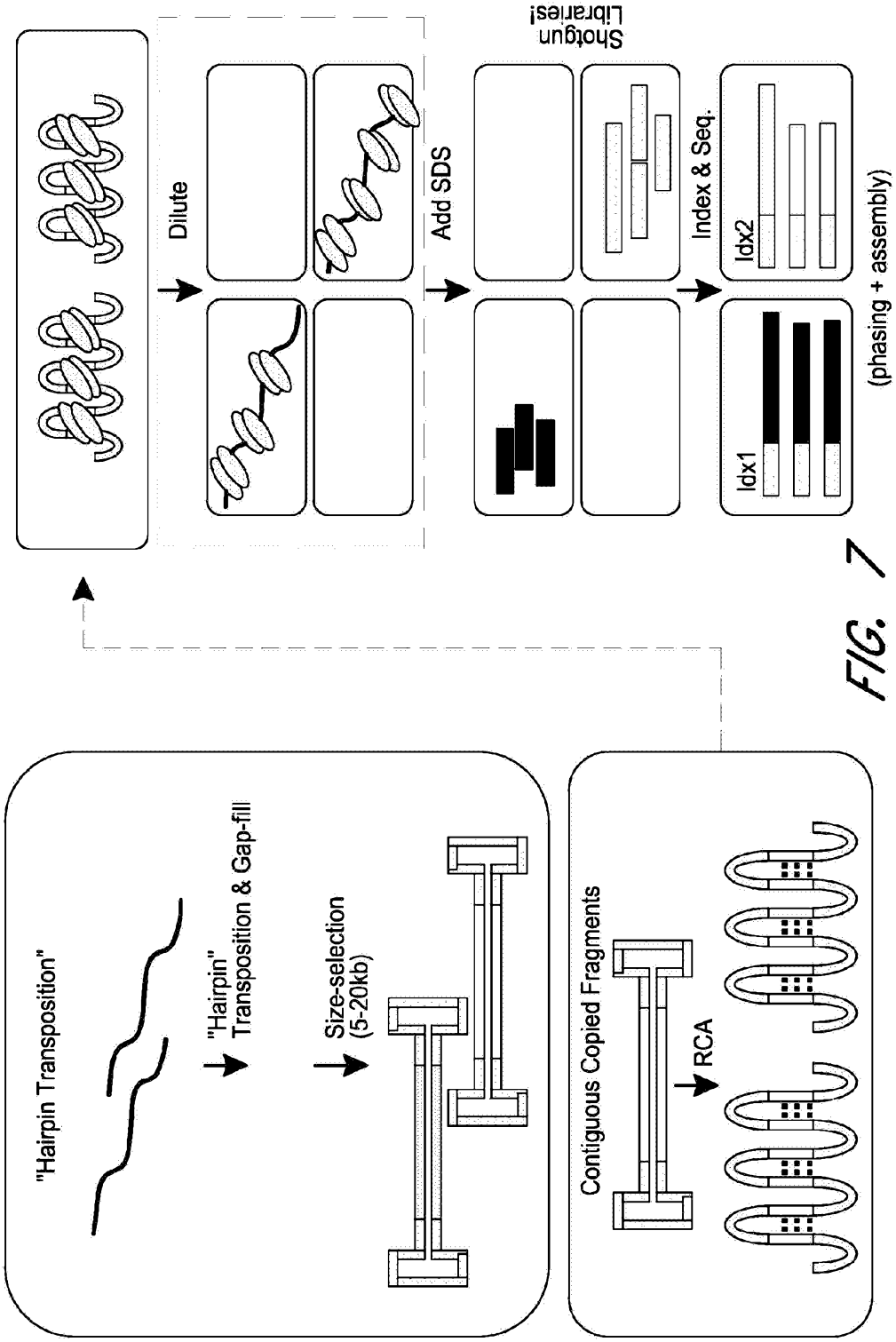

"Hairpin Transposition"
"Hairpin" Transposition & Gap-fill
Size-selection (5-20kb)
Contiguous Copied Fragments
RCA
Dilute
Add SDS
Shotgun Libraries!
Index & Seq.
Idx1
Idx2
(phasing + assembly)

FIG. 7

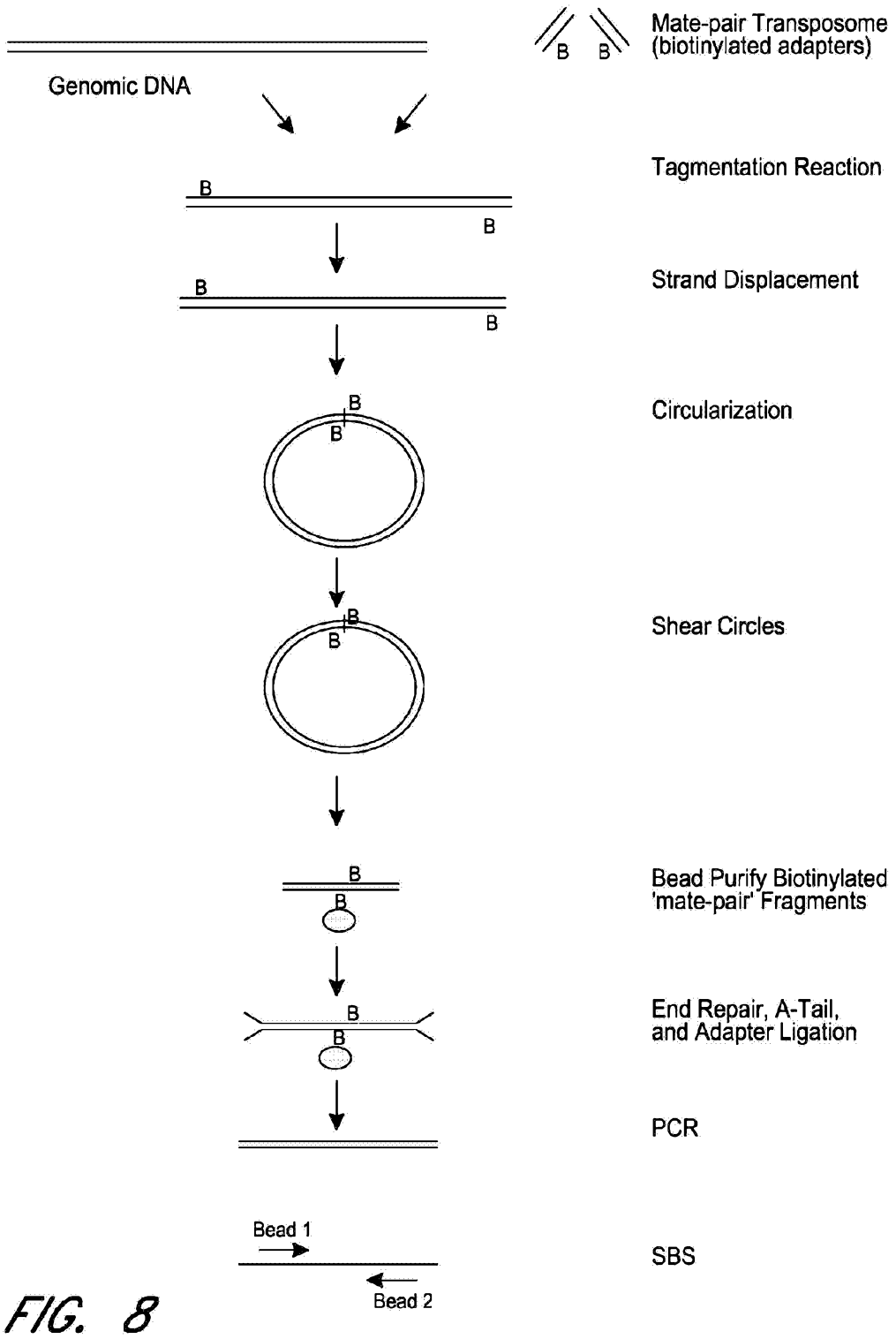

Mate-pair Transposome
(biotinylated adapters)
B
B
Genomic DNA
Tagmentation Reaction
B
B
Strand Displacement
B
B
Circularization
B
B
Shear Circles
B
B
Bead Purify Biotinylated
'mate-pair' Fragments
B
B
End Repair, A-Tail,
and Adapter Ligation
PCR
Bead 1
Bead 2
SBS

FIG. 8

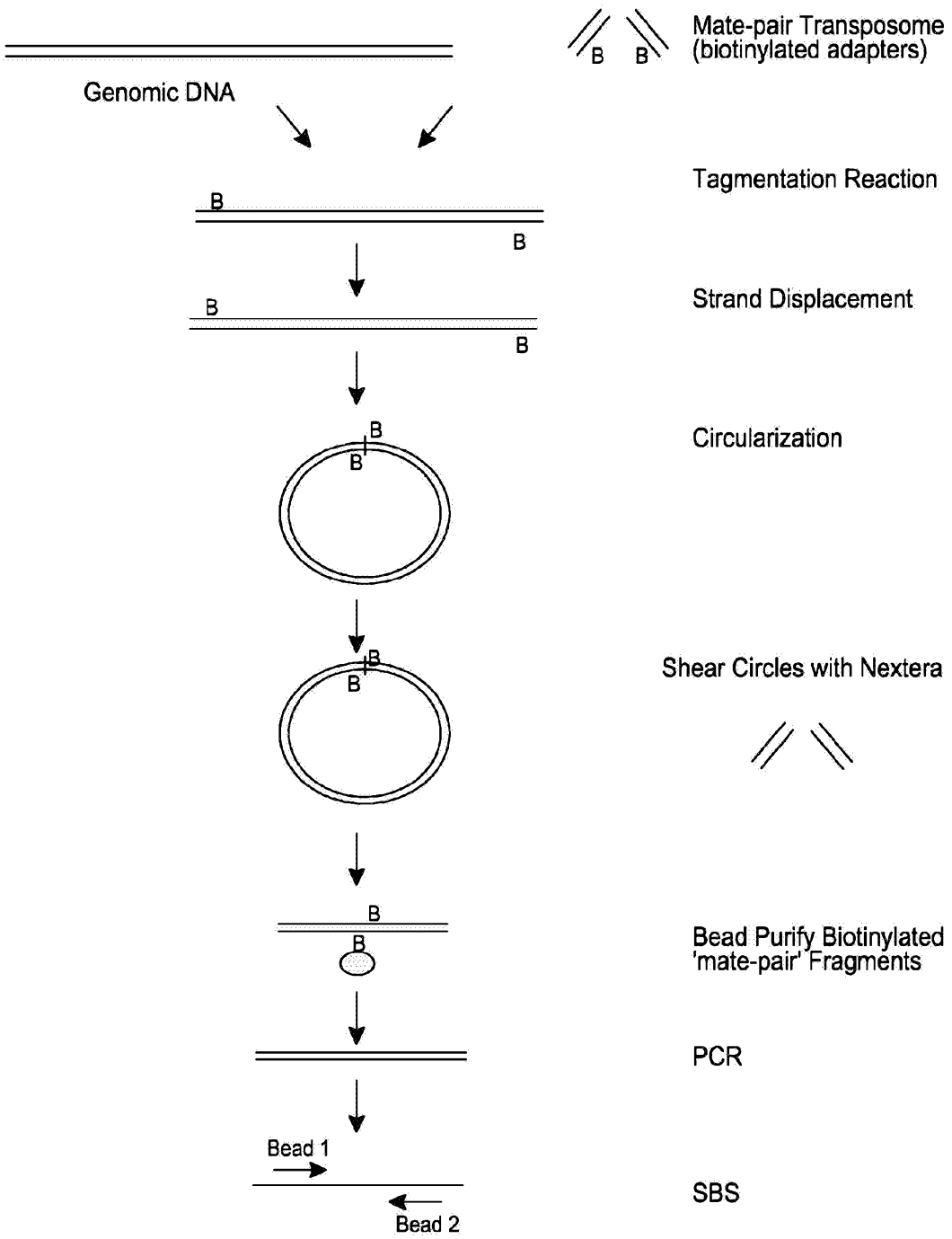
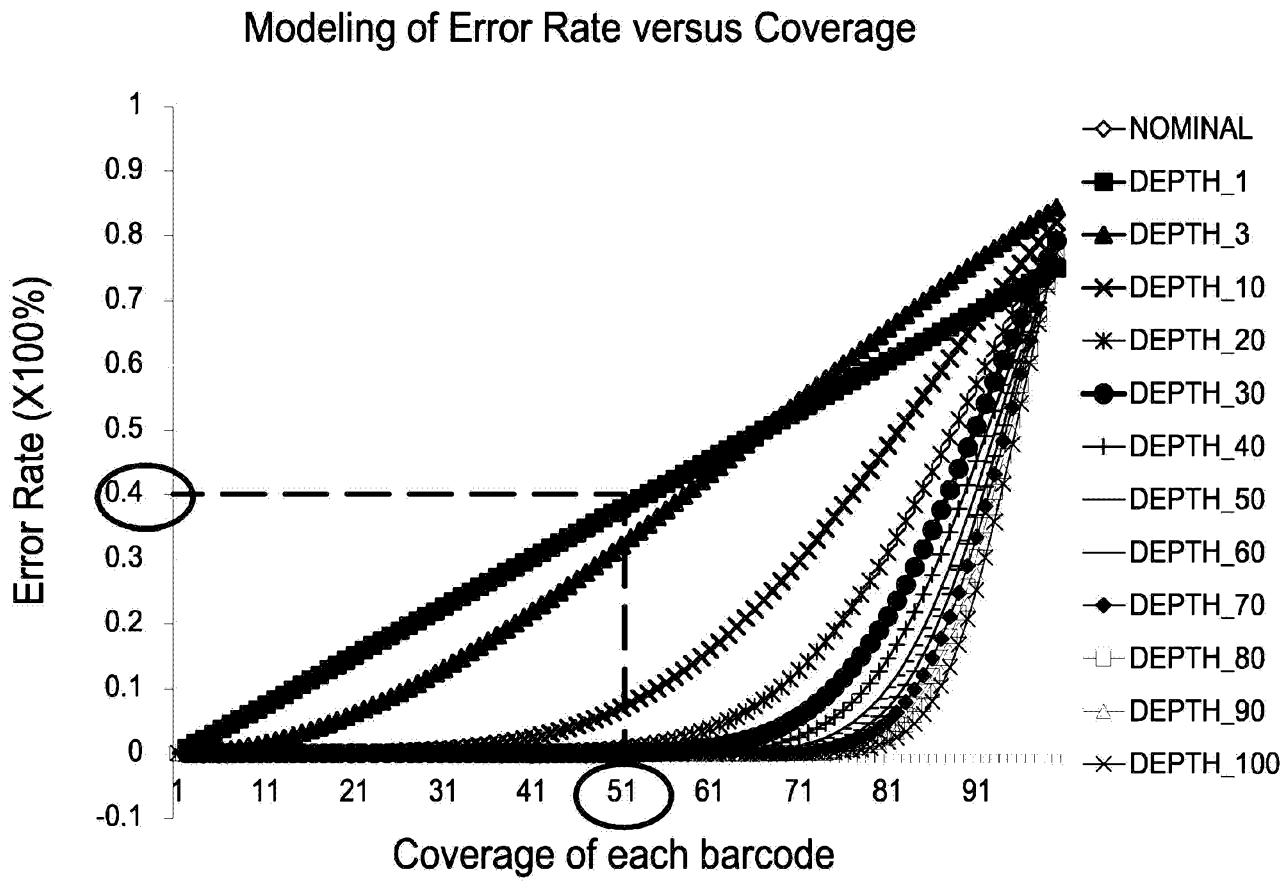

Mate-pair Transposome
(biotinylated adapters)
B
B
Genomic DNA
Tagmentation Reaction
B
B
Strand Displacement
B
B
Circularization
B
B
Shear Circles with Nextera
B
B
Bead Purify Biotinylated
'mate-pair' Fragments
B
B
PCR
Bead 1
SBS
Bead 2

FIG. 9



- At 10X coverage, error rate of 0.4 is reduced to 0.09
- At 50X coverage, error rate of 0.4 is reduced to 4.1×10^{-5}

FIG. 10

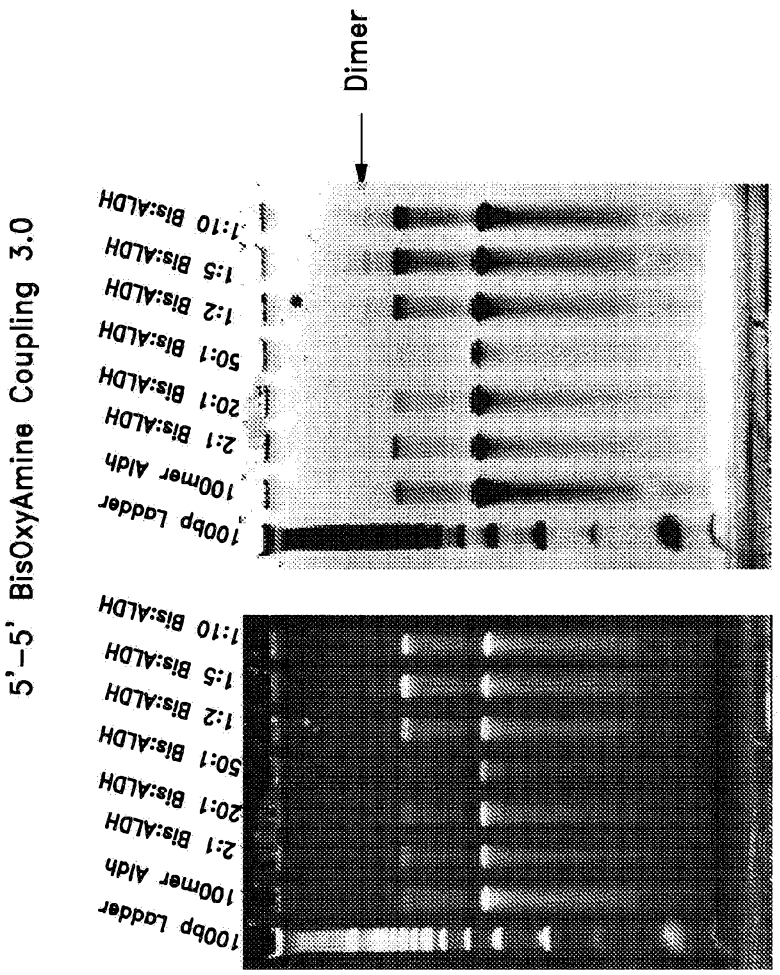

5'-5' BisOxyAmine Coupling 3.0
Dimer
100bp Ladder
100mer Aldh
2:1 Bis:ALDH
20:1 Bis:ALDH
50:1 Bis:ALDH
1:2 Bis:ALDH
1:5 Bis:ALDH
1:10 Bis:ALDH
100bp Ladder
100mer Aldh
2:1 Bis:ALDH
20:1 Bis:ALDH
50:1 Bis:ALDH
1:2 Bis:ALDH
1:5 Bis:ALDH
1:10 Bis:ALDH

FIG. 11

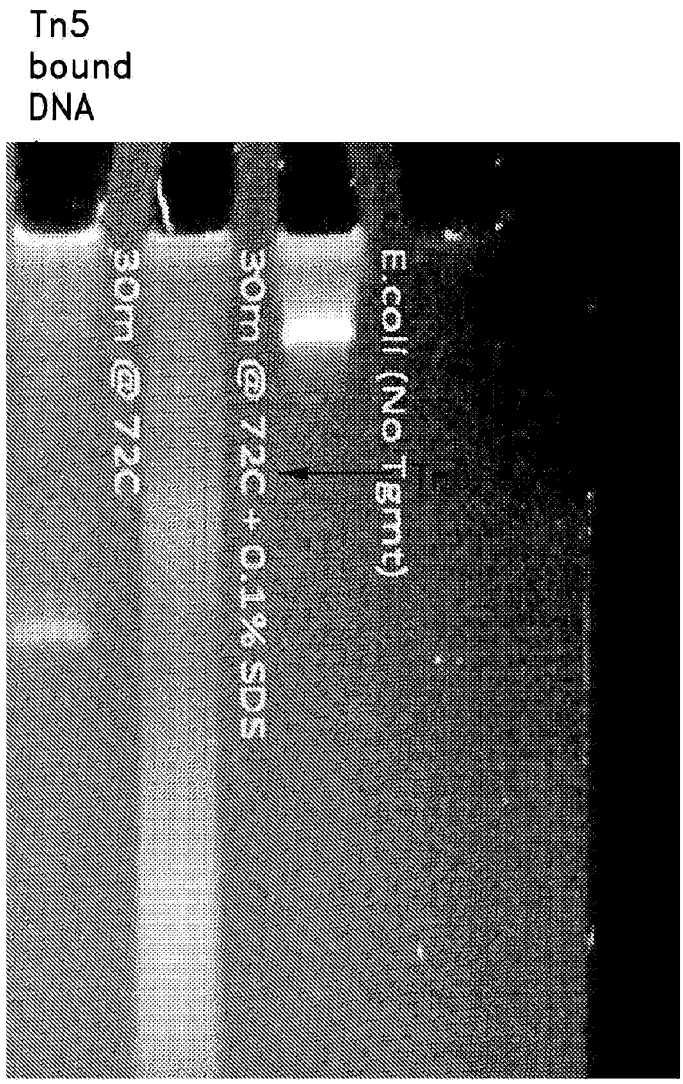


FIG. 12

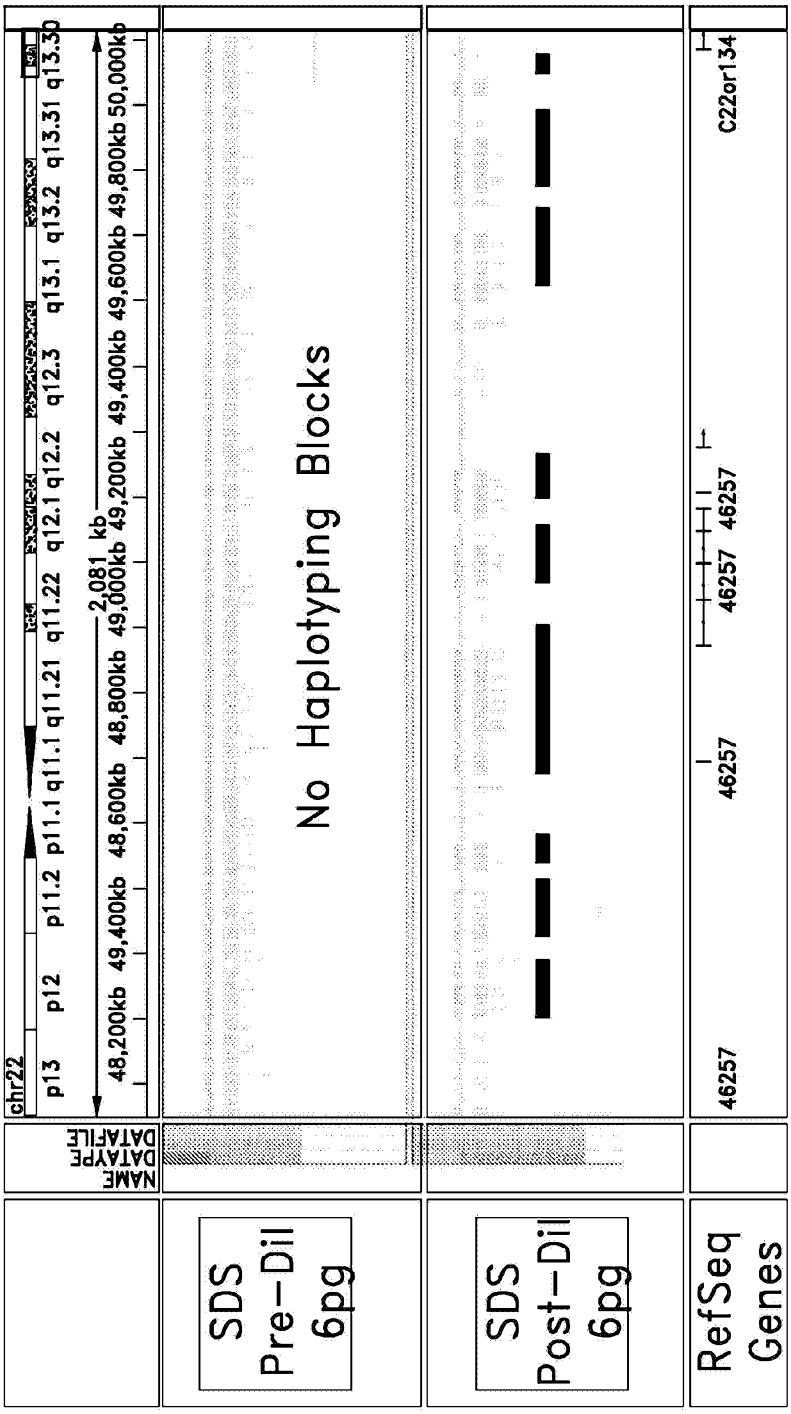

chr22
p13
p12
p11.2
p11.1
q11.1
q11.21
q11.22
q12.1
q12.2
q12.3
q13.1
q13.2
q13.31
q13.3b
48,200kb
49,400kb
48,600kb
48,800kb
2,081 kb
49,000kb
49,200kb
49,400kb
49,600kb
49,800kb
50,000kb
NAME
DATATYPE
DATAFILE
SDS Pre-Dil 6pg
No Haplotyping Blocks
SDS Post-Dil 6pg
RefSeq Genes
46257
46257
46257
46257
46257
C22or134

FIG. 13

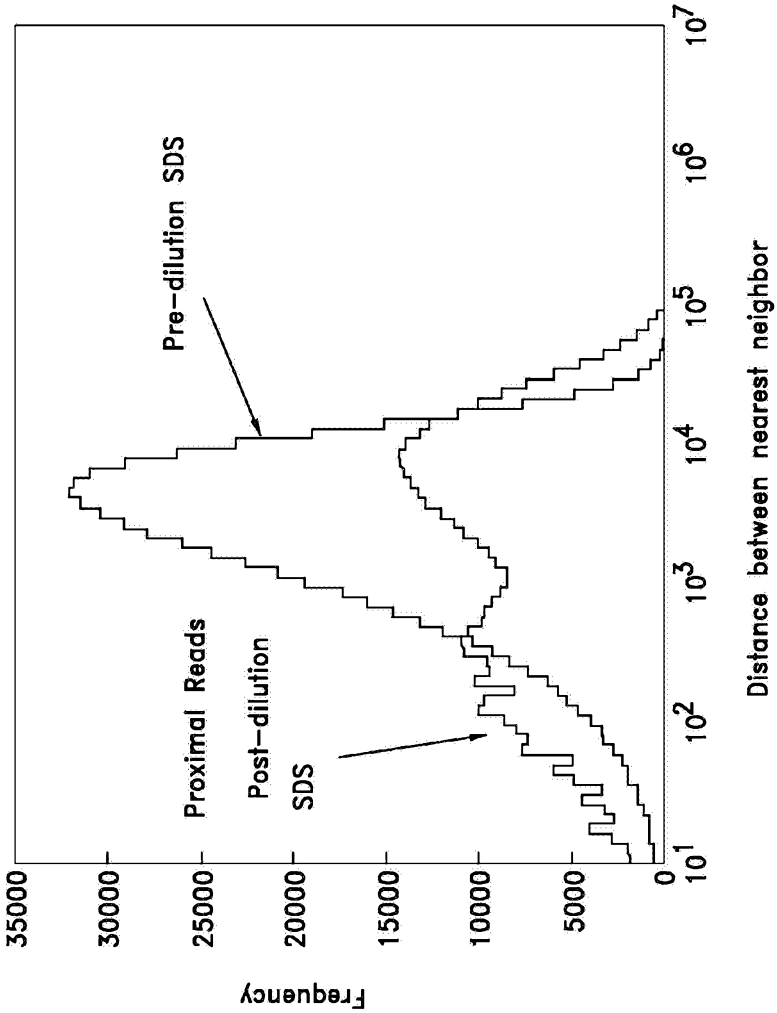

Frequency
35000
30000
25000
20000
15000
10000
5000
0
Proximal Reads
Post-dilution SDS
Pre-dilution SDS
10^1
10^2
10^3
10^4
10^5
10^6
10^7
Distance between nearest neighbor

FIG. 14

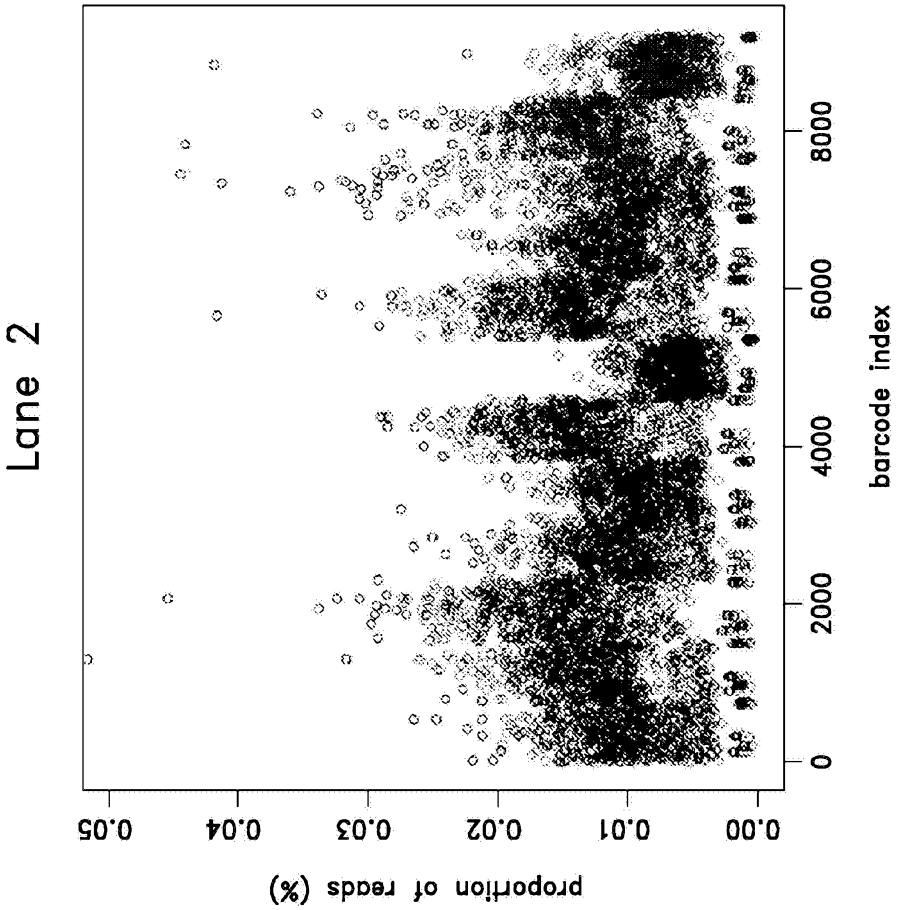

Lane 2
proportion of reads (%)
0.00
0.01
0.02
0.03
0.04
0.05
0
2000
4000
6000
8000
barcode index

FIG. 15

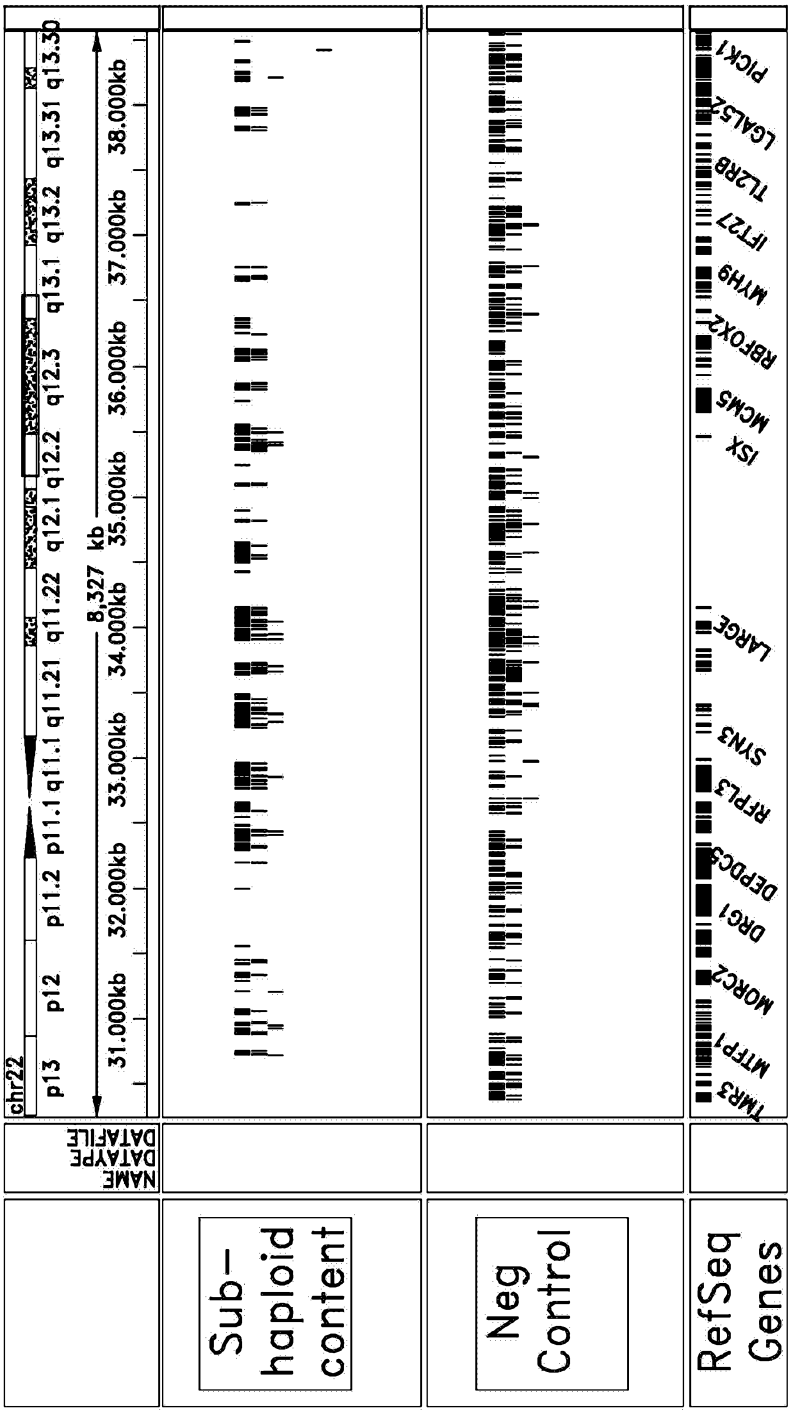

NAME
DATATYPE
DATAFILE
chr22
p13
p12
p11.2
p11.1
q11.1
q11.21
q11.22
q12.1
q12.2
q12.3
q13.1
q13.2
q13.31
q13.3b
31.000kb
32.000kb
33.000kb
34.000kb
8,327 kb
35.000kb
36.000kb
37.000kb
38.000kb
Sub-haploid content
Neg Control
RefSeq Genes
TMR3
MTFP1
MORC2
DRG1
DEPDC5
RFFL3
SYN3
LARGE
ISX
MCM5
RBFOX2
MYH9
IFT27
TL2RB
LGALS2
PICK1

FIG. 16