US 20090111705A1

(19) **United States**

(12) **Patent Application Publication**

Sparks et al.

(10) **Pub. No.: US 2009/0111705 A1**

(43) **Pub. Date: Apr. 30, 2009**

(54) **SELECTION OF DNA ADAPTOR ORIENTATION BY HYBRID CAPTURE**

(75) Inventors: **Andrew Sparks**, Los Gatos, CA (US); **Arnold Oliphant**, Sunnyvale, CA (US); **George Yeung**, Mountain View, CA (US)

Correspondence Address:
MORGAN, LEWIS & BOCKIUS, LLP
ONE MARKET SPEAR STREET TOWER
SAN FRANCISCO, CA 94105 (US)

(73) Assignee: **Complete Genomics, Inc.**, Mountain View, CA (US)

(21) Appl. No.: **11/934,697**

(22) Filed: **Nov. 2, 2007**

Related U.S. Application Data

(60) Provisional application No. 60/864,992, filed on Nov. 9, 2006.

Publication Classification

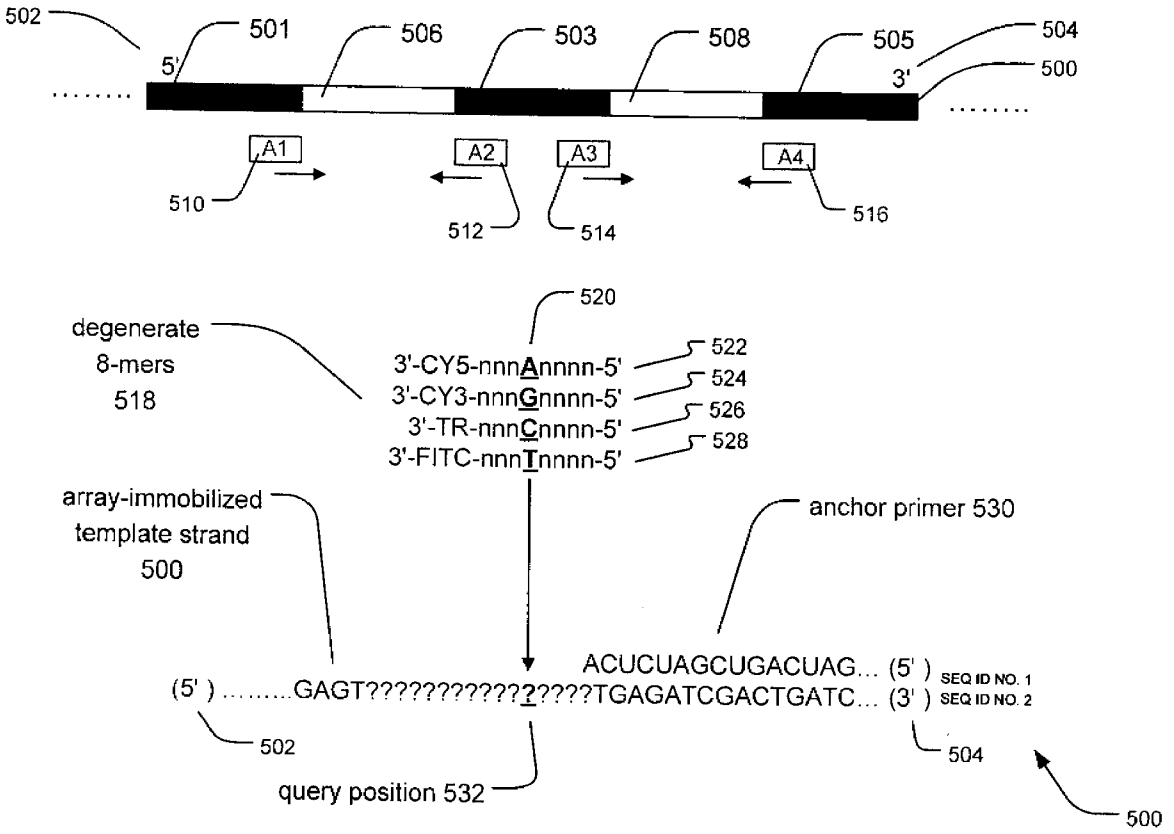
(51) **Int. Cl.**

C40B 30/04	(2006.01)
C12N 15/00	(2006.01)
C40B 40/08	(2006.01)
C12N 9/00	(2006.01)

(52) **U.S. Cl.** **506/9**; 435/320.1; 506/17; 435/183

(57) **ABSTRACT**

Aspects described and claimed herein provide methods to insert multiple DNA adaptors into a population of circular target DNAs at defined positions and orientations with respect to one another by employing selective capture of defined molecules. The resulting multi-adaptor constructs are then used in massively-parallel nucleic acid sequencing techniques.



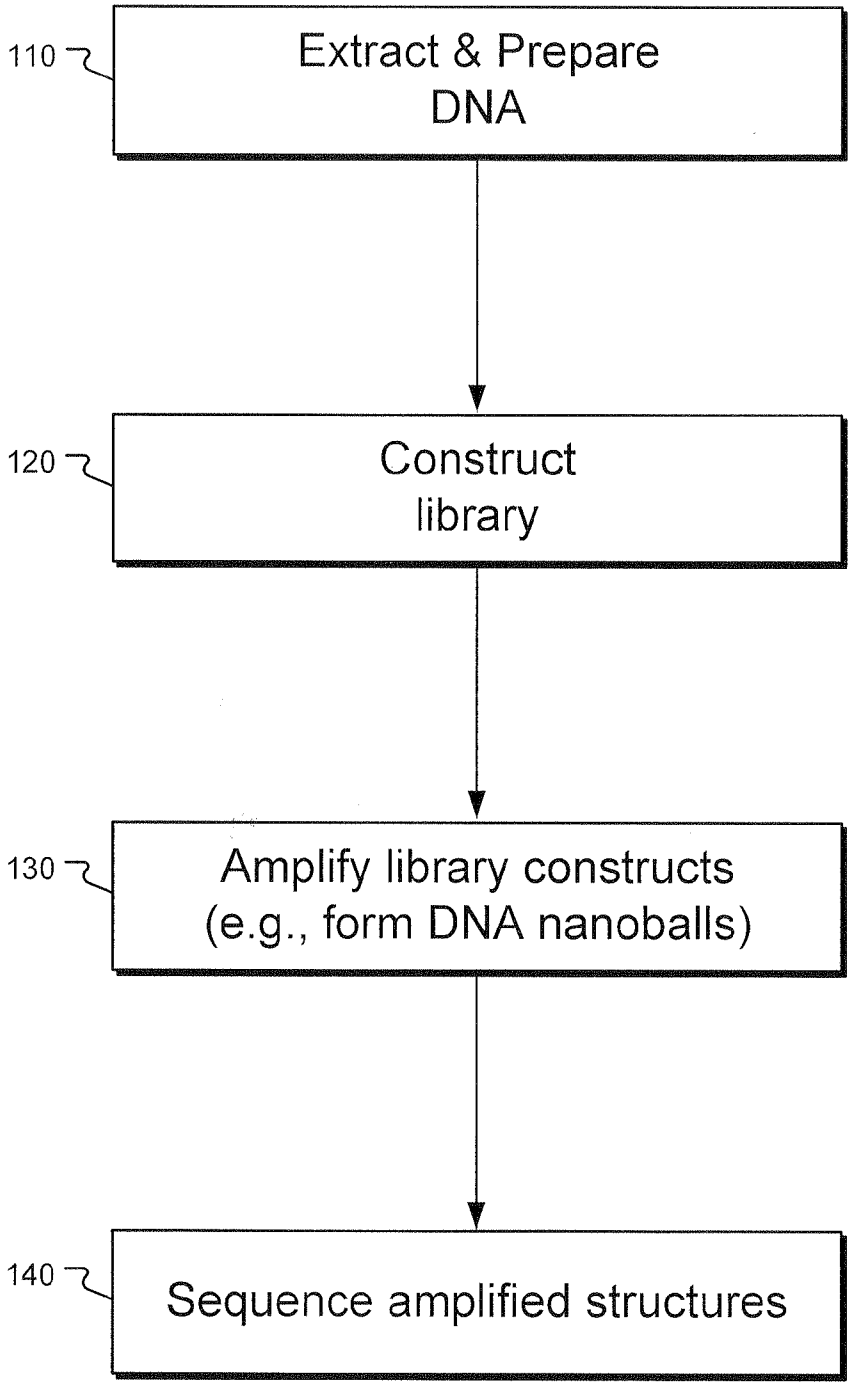

110
Extract & Prepare DNA
120
Construct library
130
Amplify library constructs (e.g., form DNA nanoballs)
140
Sequence amplified structures
100

Fig. 1

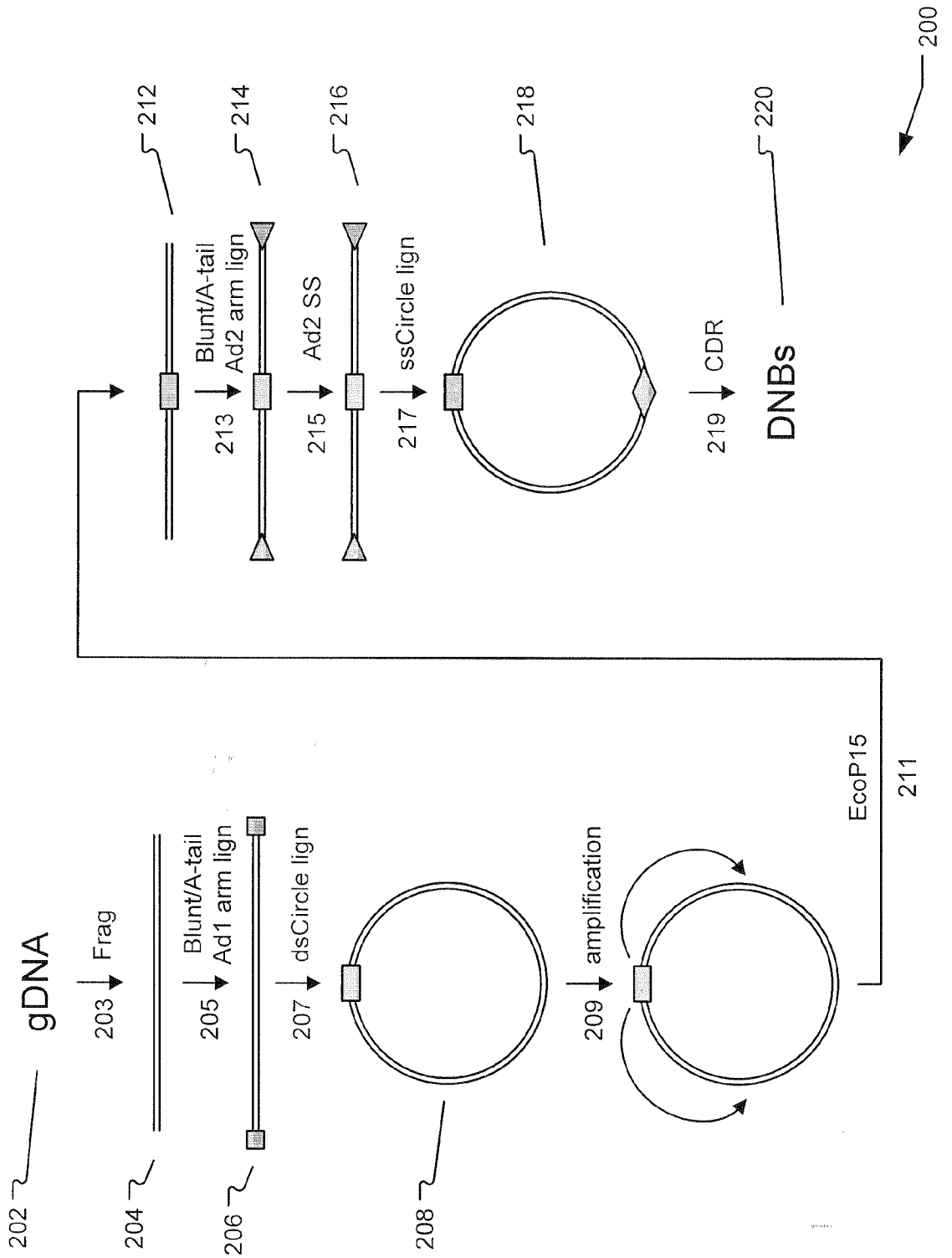

gDNA
202
Frag
203
204
Blunt/A-tail
Ad1 arm lign
205
206
dsCircle lign
207
208
amplification
209
EcoP15
211
212
Blunt/A-tail
Ad2 arm lign
213
214
Ad2 SS
215
216
ssCircle lign
217
218
CDR
219
DNBs
220
200

Fig. 2

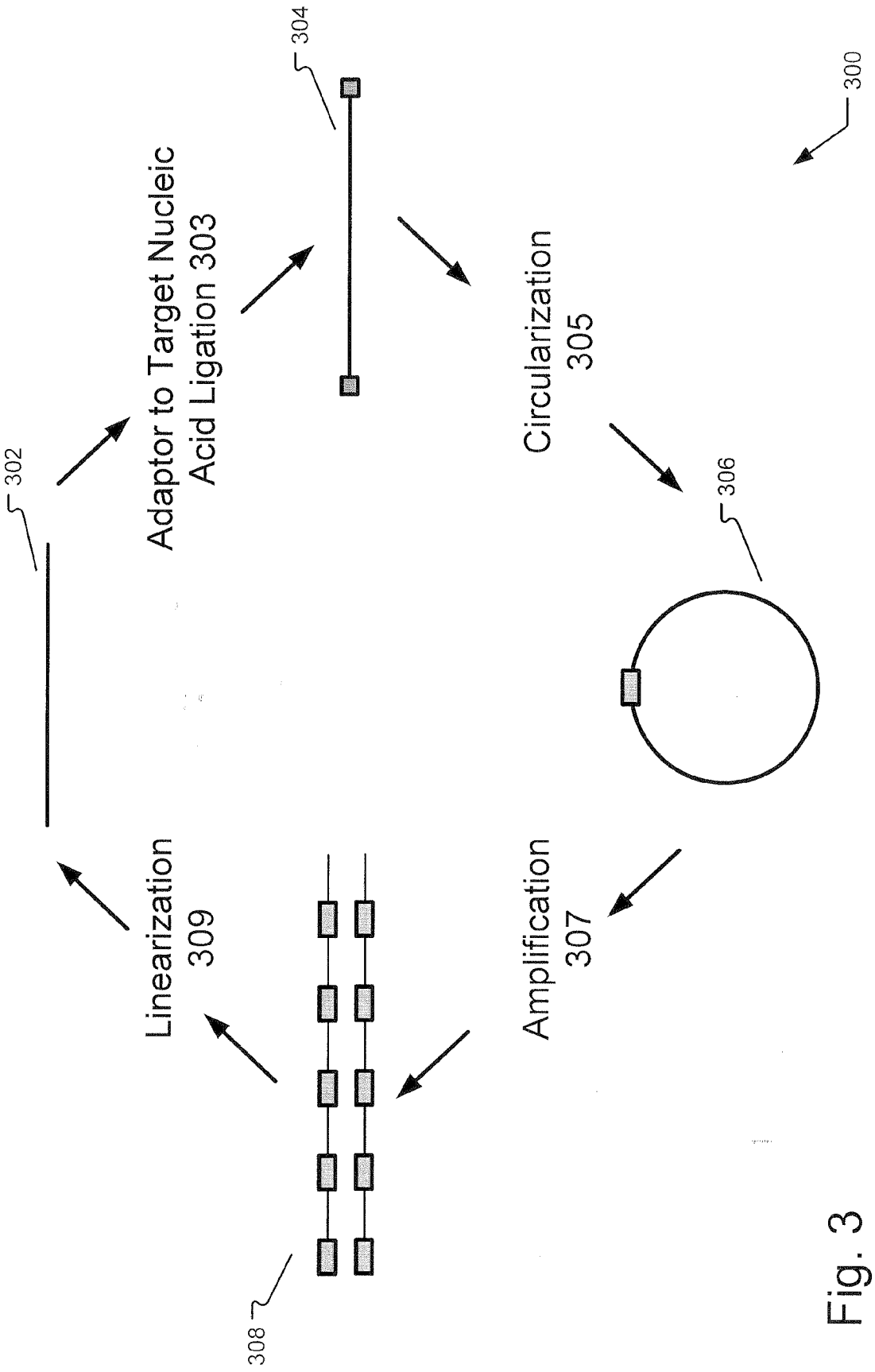

300
302
304
306
308
Adaptor to Target Nucleic Acid Ligation 303
Circularization 305
Amplification 307
Linearization 309

Fig. 3

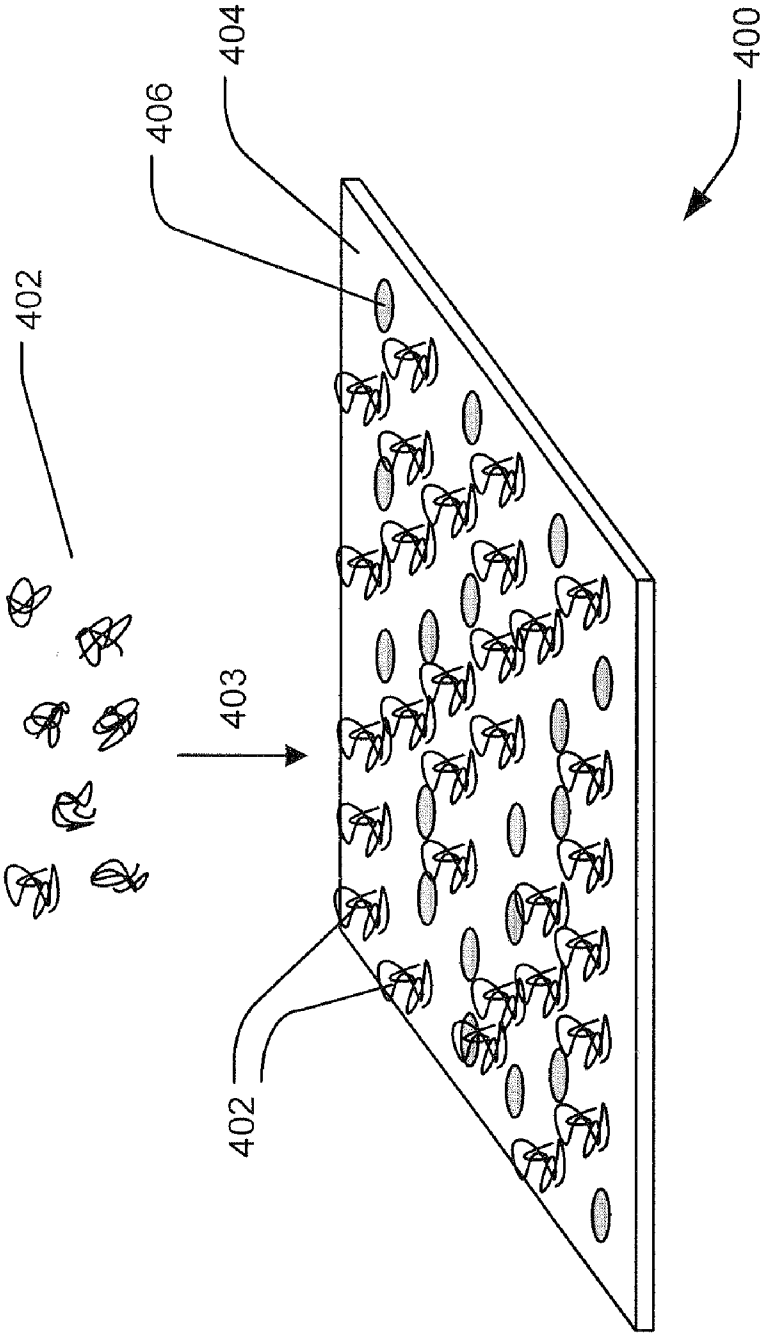


Fig. 4

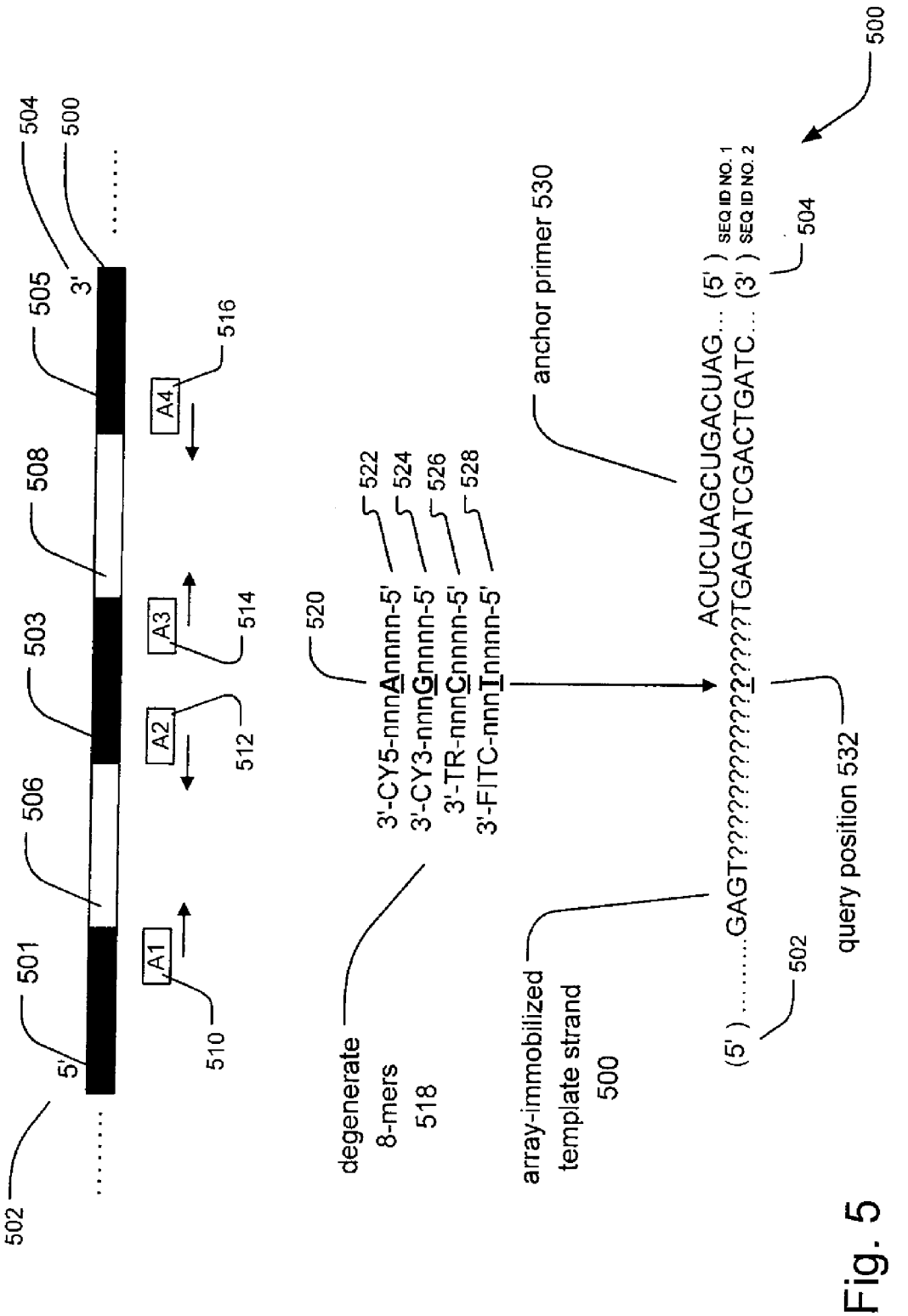

502
5'
501
506
503
508
505
3'
504
500
A1
510
A2
512
A3
514
A4
516
degenerate
8-mers
518
520
3'-CY5-nnnAnnnn-5'
3'-CY3-nnnGnnnn-5'
3'-TR-nnnCnnnn-5'
3'-FITC-nnnTnnnn-5'
522
524
526
528
anchor primer 530
array-immobilized
template strand
500
(5') .....GAGT????????????????TGAGATCGACTGATC... (3')
ACUCUAGCUGACUAG... (5') SEQ ID NO. 1
(3') SEQ ID NO. 2
502
query position 532
504
500

Fig. 5

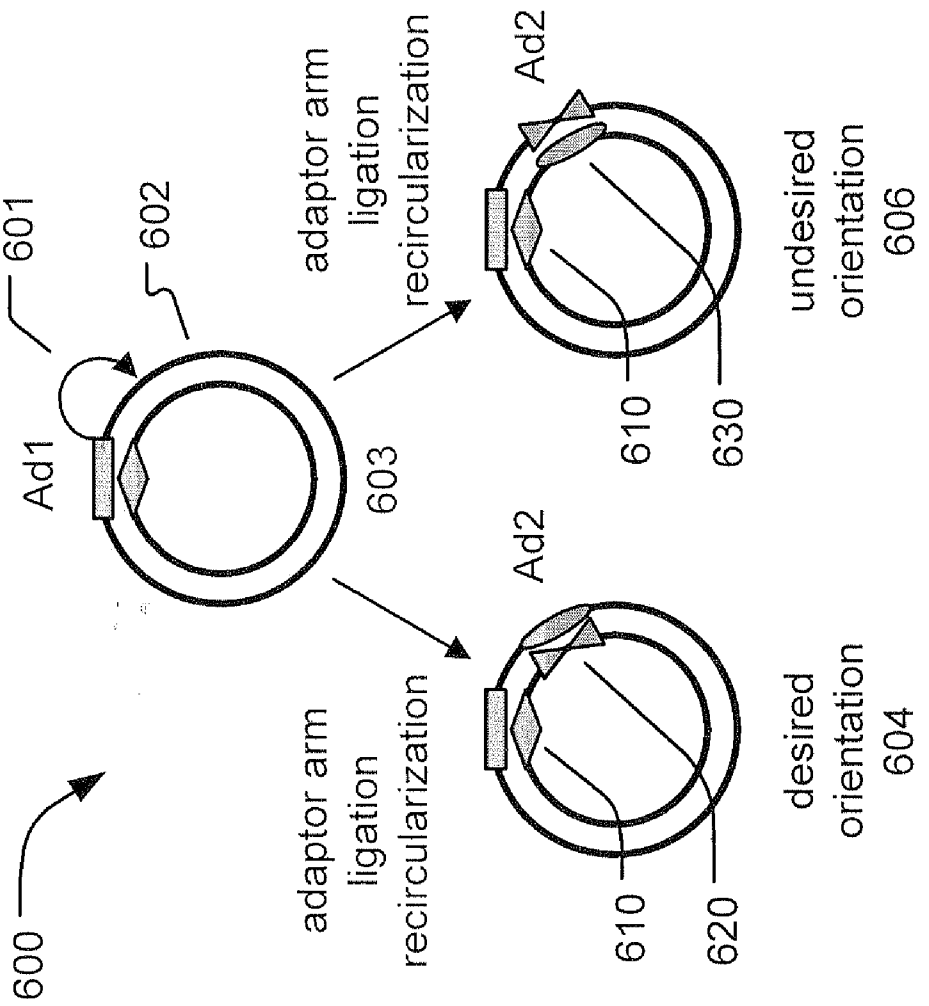

600
601
602
Ad1
603
adaptor arm ligation recircularization
adaptor arm ligation recircularization
Ad2
610
620
desired orientation 604
Ad2
610
630
undesired orientation 606

Fig. 6

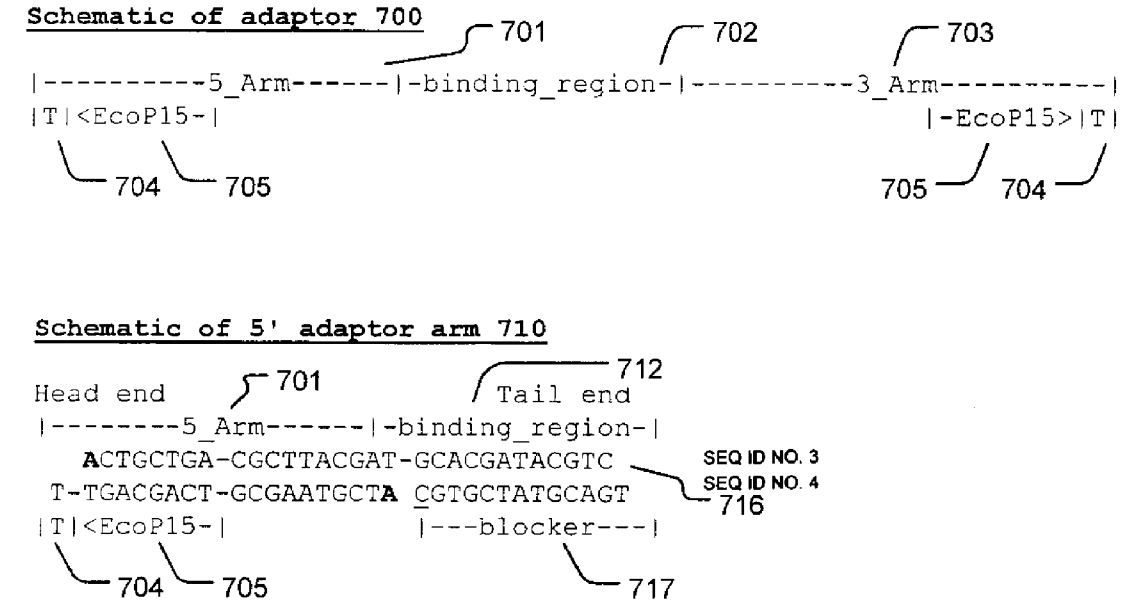

Schematic of adaptor 700
701
702
703
----------5_Arm------	-binding_region-	----------3_Arm----------
T	<EcoP15-
-EcoP15>	T
704
705
705
704
Schematic of 5' adaptor arm 710
Head end
701
712
Tail end
--------5_Arm------	-binding_region-
ACTGCTGA-CGCTTACGAT-GCACGATACGTC
SEQ ID NO. 3
T-TGACGACT-GCGAATGCTA CGTGCTATGCAGT
SEQ ID NO. 4
716
T	<EcoP15-
---blocker---
704
705
717
Schematic of 3' adaptor arm 720
704
727
705
---blocker---
-EcoP15>	T
TGCACGATACGTC TACGATGCGA-ACAGCAGA-T
SEQ ID NO. 5
CGTGCTATGCAG-ATGCTACGCT-TGTCGTCT
SEQ ID NO. 6
-binding_region-	-------3_Arm-------
tail end
Head end
726
722
703
Schematic of final adaptor 730
413333331 2 2 133333314
AACTGCTGANNNNNNNNNNGNNNNNNNNNNCNNNNNNNNNNACAGCAGAT
SEQ ID NO. 7
AACTGCTGACGCTTACGATGCACGATACGTCTACGATGCGAACAGCAGA
SEQ ID NO. 8
TGACGACTGCGAATGCTACGTGCTATGCAGATGCTACGCTTGTCGTCTA
SEQ ID NO. 9

Fig. 7

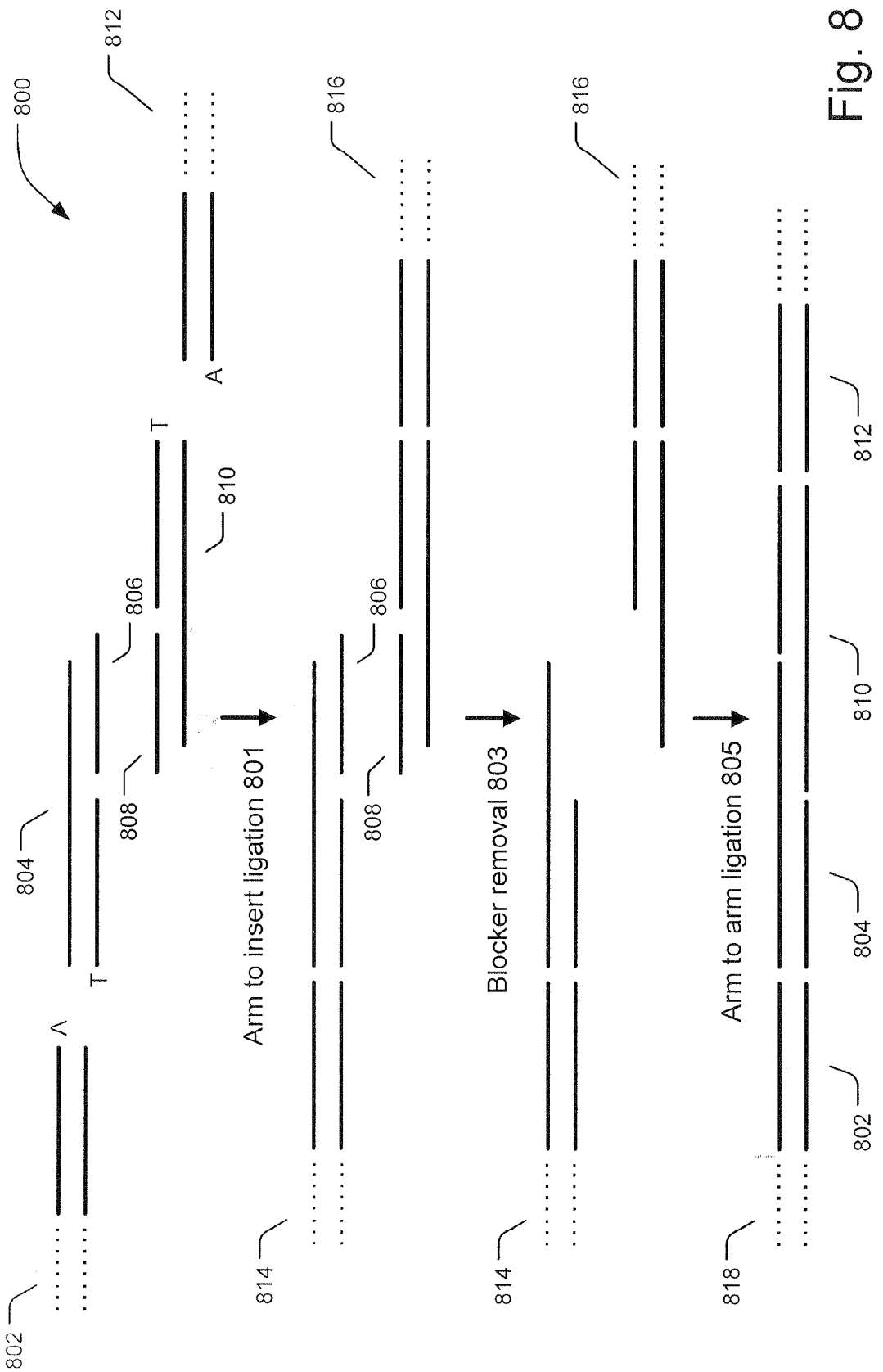

800
812
816
816
802
A
T
804
808
806
T
A
810
Arm to insert ligation 801
814
808
806
Blocker removal 803
814
Arm to arm ligation 805
818
802
804
810
812

Fig. 8

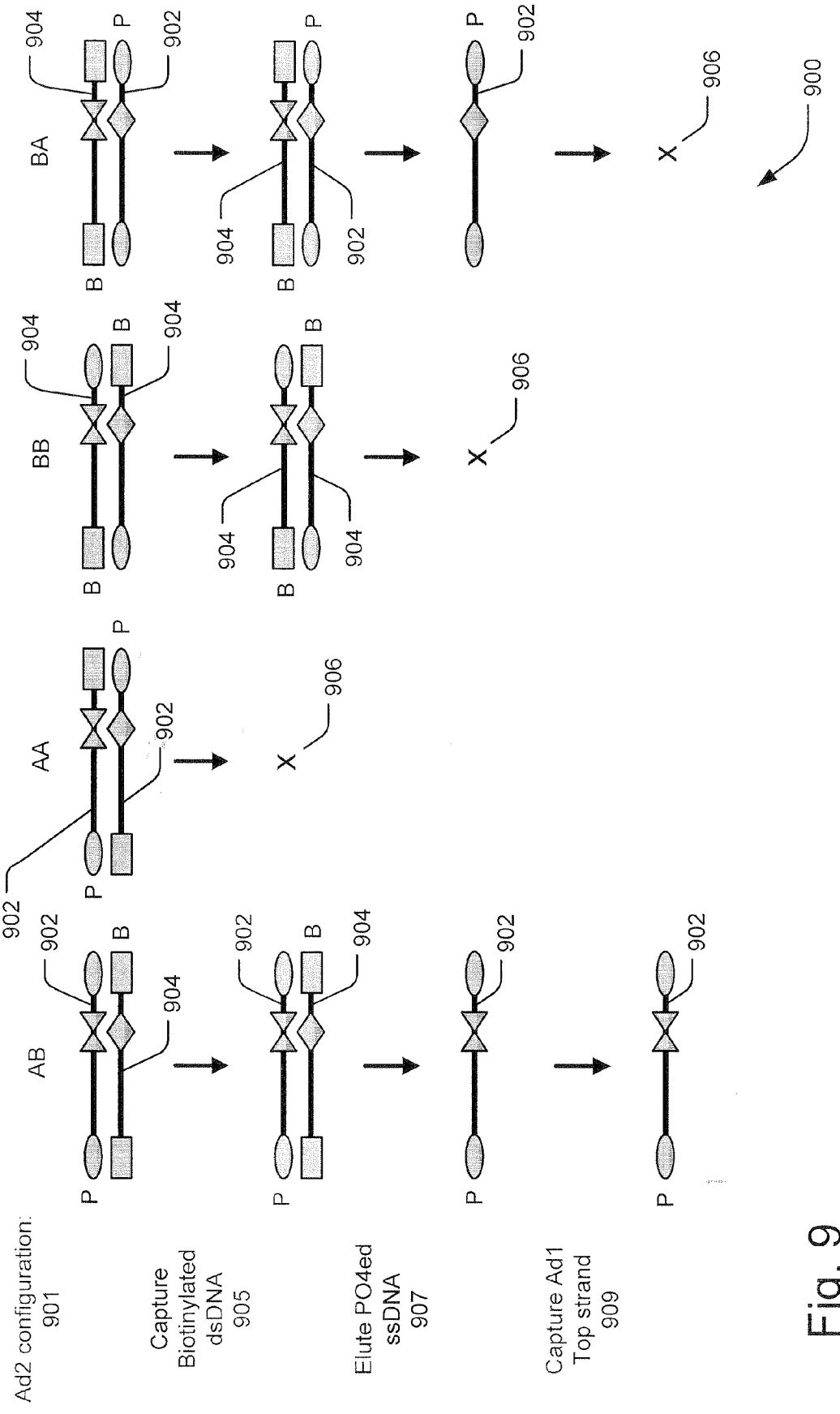

Ad2 configuration:
901
AB
AA
BB
BA
902
904
906
900
P
B
X
Capture
Biotinylated
dsDNA
905
Elute PO4ed
ssDNA
907
Capture Ad1
Top strand
909

Fig. 9

SELECTION OF DNA ADAPTOR ORIENTATION BY HYBRID CAPTURE

[0001] This application claims priority to U.S. Provisional Application 60/864,992 filed Nov. 9, 2006.

BACKGROUND

[0002] Large-scale sequence analysis of genomic DNA is central to understanding a wide range of biological phenomena related to health and disease in humans and is economically important plants and animals. The need for low-cost, high-throughput sequencing and re-sequencing has led to the development of new approaches to sequencing that employ parallel analysis of many target DNA fragments simultaneously. Improvements to sequencing methods and increasing the amount and quality of data from such methods is of great value in the art.

SUMMARY

[0003] Embodiments described and claimed herein address the foregoing and other situations by providing methods to provide repeated cycles of nucleic acid cleavage and ligation to insert multiple DNA adaptors into a population of circular target DNAs at defined positions and orientations with respect to one another. The resulting multi-adaptor constructs are then used in massively-parallel nucleic acid sequencing techniques.

[0004] Aspects of the technology provide methods for selecting for orientation of two adaptors with respect to one another in library constructs comprising: obtaining target nucleic acid; ligating a first adaptor to the target nucleic acid to produce first library constructs wherein one strand of the first adaptor comprises a capture sequence; ligating first and second arms of a second adaptor to the linearized first library constructs to form second library constructs, wherein wherein at least one strand of the second adaptor arm comprises a functional group; capturing functionalized double-stranded second library constructs and discarding un-functionalized second library constructs; denaturing and eluting single-stranded nucleic acids from the captured double-stranded functionalized second library constructs; and capturing the capture sequence in the one strand of the first adaptor, thereby selecting for orientation of the second adaptor with respect to the first adaptor in the library constructs.

[0005] Other aspects of the technology provide methods for selecting for orientation of two adaptors with respect to one another in library constructs comprising: obtaining target nucleic acid; ligating a first adaptor to the target nucleic acid to produce first library constructs; ligating first and second arms of a second adaptor to the linearized first library constructs to form second library constructs; amplifying the second library construct with a functionalized primer complementary to an end of one strand of the second adaptor; capturing functionalized amplified double-stranded second library constructs and discarding un- functionalized second library constructs; denaturing and eluting single-stranded nucleic acids from the captured double-stranded functionalized second library constructs; and capturing the capture sequence in the one strand of the first adaptor, thereby selecting for orientation of the second adaptor with respect to the first adaptor in the library constructs.

[0006] In some aspects of this method, the first library constructs are circularized between the two ligating steps. In other aspects, the first library constructs are cut with a restriction endonuclease after being circularized. In yet other aspects, the first adaptor is ligated to the target nucleic acid as two adaptor arms. Also, in other aspects the first and second adaptors further comprise Type IIs endonuclease recognition sites.

[0007] In other aspects, two-component binding systems with high affinity and specificity (e.g., avidin-biotin and antibody-hapten systems) can be used in the immobilization and isolation of certain constructs. One functional group of the pair can be attached to an adaptor for use in the isolation of molecules comprising that adaptor. For example, the functional group can be biotin and the functionalized double-stranded second library constructs can be captured by strepavidin.

[0008] Another aspect of the invention provides a method for selecting for orientation of two or more adaptors with respect to one another in library constructs comprising: (a) obtaining target nucleic acid; (b) ligating a first adaptor to the target nucleic acid to produce first library constructs wherein one strand of the first adaptor comprises a capture sequence; (c) ligating first and second arms of a second adaptor to the linearized first library construct to form second library constructs, wherein at least one end of the second adaptor arm comprises; (d) capturing functionalized double-stranded second library constructs and discarding un-functionalized second library constructs; (e) denaturing and eluting single-stranded nucleic acids from the captured double-stranded functionalized second library constructs; (f) capturing the capture sequence in the one strand of the first adaptor, thereby selecting for orientation of the second adaptor to the first adaptor in the library constructs; (g) repeating processes (b) through (f) until a desired number of adaptors have been inserted into the nucleic acid library constructs.

[0009] In some aspects, either the first or second adaptor arms of the second adaptor that is ligated to the first library construct comprises a functional group; in other aspects, after ligation of the first and second arms of the second adaptor, the second library constructs are amplified with a functionalized primer complementary to an end of one strand of the second adaptor, and the functionalized amplified double-stranded second library constructs are captured.

[0010] Also in some aspects, amplicons made by sequentially selectively capturing a functionalized adaptor and selectively capturing one strand of one adaptor in a library construct are provided, as are libraries comprising a multiplicity (five or more) of such amplicons. In other aspects, kits are provided for selecting for desired orientations of multiple adaptors in library constructs comprising a first double-stranded adaptor; a functionalized second double stranded adaptor; reagents for capturing the functionalized second double stranded adaptor; and reagents for capturing one strand of the first double-stranded adaptor.

[0011] This Summary is provided to introduce a selection of concepts in a simplified form that are further described below in the Detailed Description. This Summary is not intended to identify key or essential features of the claimed subject matter, nor is it intended to be used to limit the scope of the claimed subject matter. Other features, details, utilities, and advantages of the claimed subject matter will be apparent from the following written Detailed Description including

those aspects illustrated in the accompanying drawings and defined in the appended claims.

BRIEF DESCRIPTIONS OF THE DRAWINGS

[0012] FIG. **1** is a simplified flow diagram of an overall method for sequencing nucleic acids using the processes of the claimed invention.

[0013] FIG. **2** is a schematic representation of one aspect of a method for assembling adaptor/target nucleic acid library constructs.

[0014] FIG. **3** is a schematic illustration of a basic adaptor insertion process.

[0015] FIG. **4** is a schematic illustration of one aspect of a DNA array employing multi-adaptor nucleic acid library constructs.

[0016] FIG. **5** is a schematic illustration of the components that may be used in an exemplary sequencing-by-ligation technique.

[0017] FIG. **6** is a schematic illustration of an insertion of a second adaptor relative to a first adaptor in a nucleic acid library construct.

[0018] FIG. **7** is a schematic representation of components of an exemplary adaptor useful for selecting insertion orientation.

[0019] FIG. **8** is a schematic representation of adaptor insertion allowing subsequent circularization of the target/adaptor construct.

[0020] FIG. **9** is a schematic illustration of a hybrid capture process for selecting constructs where adaptors are inserted into a target nucleic acid in a desired orientation.

DEFINITIONS

[0021] The practice of the techniques described herein may employ, unless otherwise indicated, conventional techniques and descriptions of organic chemistry, polymer technology, molecular biology (including recombinant techniques), cell biology, biochemistry, and sequencing technology, which are within the skill of those who practice in the art. Such conventional techniques include polymer array synthesis, hybridization and ligation of polynucleotides, and detection of hybridization using a label. Specific illustrations of suitable techniques can be had by reference to the examples herein. However, other equivalent conventional procedures can, of course, also be used. Such conventional techniques and descriptions can be found in standard laboratory manuals such as Green, et al., Eds. (1999), *Genome Analysis: A Laboratory Manual Series* (Vols. I-IV); Weiner, Gabriel, Stephens, Eds. (2007), *Genetic Variation: A Laboratory Manual*; Dieffenbach, Dveksler, Eds. (2003), *PCR Primer: A Laboratory Manual*; Bowtell and Sambrook (2003), *DNA Microarrays: A Molecular Cloning Manual*; Mount (2004), *Bioinformatics: Sequence and Genome Analysis*; Sambrook and Russell (2006), *Condensed Protocols from Molecular Cloning: A Laboratory Manual*; and Sambrook and Russell (2002), *Molecular Cloning: A Laboratory Manual* (all from Cold Spring Harbor Laboratory Press); Stryer, L. (1995) *Biochemistry* (4th Ed.) W.H. Freeman, New York N.Y.; Gait, "*Oligonucleotide Synthesis: A Practical Approach*" 1984, IRL Press, London; Nelson and Cox (2000), *Lehninger, Principles of Biochemistry* 3rd Ed., W.H. Freeman Pub., New York, N.Y.; and Berg et al. (2002) *Biochemistry,* 5th Ed., W.H. Freeman Pub., New York, N.Y., all of which are herein incorporated in their entirety by reference for all purposes.

[0022] Note that as used herein and in the appended claims, the singular forms "a," "an," and "the" include plural referents unless the context clearly dictates otherwise. Thus, for example, reference to "an agent" refers to one agent or mixtures of agents, and reference to "the method of administration" includes reference to equivalent steps and methods known to those skilled in the art, and so forth.

[0023] Unless defined otherwise, all technical and scientific terms used herein have the same meaning as commonly understood by one of ordinary skill in the art to which this invention belongs. All publications mentioned herein are incorporated herein by reference for the purpose of describing and disclosing devices, formulations and methodologies which are described in the publication and which might be used in connection with the presently described invention.

[0024] Where a range of values is provided, it is understood that each intervening value, between the upper and lower limit of that range and any other stated or intervening value in that stated range is encompassed within the invention. The upper and lower limits of these smaller ranges may independently be included in the smaller ranges, and are also encompassed within the invention, subject to any specifically excluded limit in the stated range. Where the stated range includes one or both of the limits, ranges excluding either both of those included limits are also included in the invention.

[0025] In the following description, numerous specific details are set forth to provide a more thorough understanding of the present invention. However, it will be apparent to one of skill in the art that the present invention may be practiced without one or more of these specific details. In other instances, well-known features and procedures well known to those skilled in the art have not been described in order to avoid obscuring the invention.

[0026] "Adaptor" refers to an engineered construct comprising "adaptor elements" where one or more adaptors may be interspersed within target nucleic acid in a library construct. The adaptor elements or features included in any adaptor vary widely depending on the use of the adaptors, but typically include sites for restriction endonuclease recognition and/or cutting, sites for primer binding (for amplifying the library constructs) or anchor primer binding (for sequencing the target nucleic acids in the library constructs), nickase sites, and the like. In some aspects, adaptors are engineered so as to comprise one or more of the following: 1) a length of about 20 to about 250 nucleotides, or about 40 to about 100 oligonucleotides, or less than about 60 nucleotides, or less than about 50 nucleotides; 2) features so as to be ligated to the target nucleic acid as two "arms"; 3) different and distinct anchor binding sites at the 5' and the 3' ends of the adaptor for use in sequencing of adjacent target nucleic acid; and 4) one or more restriction sites.

[0027] "Amplicon" means the product of a polynucleotide amplification reaction. That is, it is a population of polynucleotides that are replicated from one or more starting sequences. Amplicons may be produced by a variety of amplification reactions, including but not limited to polymerase chain reactions (PCRs), linear polymerase reactions, nucleic acid sequence-based amplification, circle dependant amplification and like reactions (see, e.g., U.S. Pat. Nos. 4,683,195; 4,965,188; 4,683,202; 4,800159; 5,210,015; 6,174,670; 5,399,491; 6,287,824 and 5,854,033; and US Pub. No. 2006/0024711).

[0028] "Circle dependant replication" or "CDR" refers to multiple displacement amplification of a double-stranded circular template using one or more primers annealing to the same strand of the circular template to generate products representing only one strand of the template. In CDR, no additional primer binding sites are generated and the amount of product increases only linearly with time. The primer(s) used may be of a random sequence (e.g., one or more random hexamers) or may have a specific sequence to select for amplification of a desired product. Without further modification of the end product, CDR often results in the creation of a linear construct having multiple copies of a strand of the circular template in tandem, i.e. a linear, single-stranded concatamer of multiple copies of a strand of the template.

[0029] "Circle dependant amplification" or "CDA" refers to multiple displacement amplification of a double-stranded circular template using primers annealing to both strands of the circular template to generate products representing both strands of the template, resulting in a cascade of multiple-hybridization, primer-extension and strand-displacement events. This leads to an exponential increase in the number of primer binding sites, with a consequent exponential increase in the amount of product generated over time. The primers used may be of a random sequence (e.g., random hexamers) or may have a specific sequence to select for amplification of a desired product. CDA results in a set of concatemeric double-stranded fragments is formed.

[0030] "Complementary" or "substantially complementary" refers to the hybridization or base pairing or the formation of a duplex between nucleotides or nucleic acids, such as, for instance, between the two strands of a double-stranded DNA molecule or between an oligonucleotide primer and a primer binding site on a single-stranded nucleic acid. Complementary nucleotides are, generally, A and T (or A and U), or C and G. Two single-stranded RNA or DNA molecules are said to be substantially complementary when the nucleotides of one strand, optimally aligned and compared and with appropriate nucleotide insertions or deletions, pair with at least about 80% of the other strand, usually at least about 90% to about 95%, and even about 98% to about 100%.

[0031] "Duplex" means at least two oligonucleotides or polynucleotides that are fully or partially complementary and which undergo Watson-Crick type base pairing among all or most of their nucleotides so that a stable complex is formed. The terms "annealing" and "hybridization" are used interchangeably to mean formation of a stable duplex. "Perfectly matched" in reference to a duplex means that the poly- or oligonucleotide strands making up the duplex form a double-stranded structure with one another such that every nucleotide in each strand undergoes Watson-Crick base pairing with a nucleotide in the other strand. A "mismatch" in a duplex between two oligonucleotides or polynucleotides means that a pair of nucleotides in the duplex fails to undergo Watson-Crick basepairing.

[0032] "Hybridization" refers to the process in which two single-stranded polynucleotides bind non-covalently to form a stable double-stranded polynucleotide. The resulting (usually) double-stranded polynucleotide is a "hybrid" or "duplex." "Hybridization conditions" will typically include salt concentrations of less than about 1M, more usually less than about 500 mM and may be less than about 200 mM. A "hybridization buffer" is a buffered salt solution such as 5% SSPE, or other such buffers known in the art. Hybridization temperatures can be as low as 5° C., but are typically greater than 22° C., and more typically greater than about 30° C., and typically in excess of 37° C. Hybridizations are usually performed under stringent conditions, i.e., conditions under which a probe will hybridize to its target subsequence but will not hybridize to the other, uncomplimentary sequences. Stringent conditions are sequence-dependent and are different in different circumstances. For example, longer fragments may require higher hybridization temperatures for specific hybridization than short fragments. As other factors may affect the stringency of hybridization, including base composition and length of the complementary strands, presence of organic solvents, and the extent of base mismatching, the combination of parameters is more important than the absolute measure of any one parameter alone. Generally stringent conditions are selected to be about 5° C. lower than the T_m for the specific sequence at a defined ionic strength and pH. Exemplary stringent conditions include a salt concentration of at least 0.01M to no more than 1M sodium ion concentration (or other salt) at a pH of about 7.0 to about 8.3 and a temperature of at least 25° C. For example, conditions of 5×SSPE (750 mM NaCl, 50 mM sodium phosphate, 5 mM EDTA at pH 7.4) and a temperature of 30° C. are suitable for allele-specific probe hybridizations.

[0033] "Ligation" means to form a covalent bond or linkage between the termini of two or more nucleic acids, e.g., oligonucleotides and/or polynucleotides, in a template-driven reaction. The nature of the bond or linkage may vary widely and the ligation may be carried out enzymatically or chemically. As used herein, ligations are usually carried out enzymatically to form a phosphodiester linkage between a 5' carbon terminal nucleotide of one oligonucleotide with a 3' carbon of another nucleotide. Template driven ligation reactions are described in the following references: U.S. Pat. Nos. 4,883,750; 5,476,930; 5,593,826; and 5,871,921.

[0034] "Microarray" or "array" refers to a solid phase support having a surface, preferably but not exclusively a planar or substantially planar surface, which carries an array of sites containing nucleic acids such that each site of the array comprises identical copies of oligonucleotides or polynucleotides and is spatially defined and not overlapping with other member sites of the array; that is, the sites are spatially discrete. The array or microarray can also comprise a non-planar interrogatable structure with a surface such as a bead or a well. The oligonucleotides or polynucleotides of the array may be covalently bound to the solid support, or may be non-covalently bound. Conventional microarray technology is reviewed in, e.g., Schena, Ed. (2000), *Microarrays: A Practical Approach* (IRL Press, Oxford). As used herein, "random array" or "random microarray" refers to a microarray where the identity of the oligonucleotides or polynucleotides is not discernable, at least initially, from their location but may be determined by a particular operation on the array, such as by sequencing, hybridizing decoding probes or the like. See, e.g., U.S. Pat. Nos. 6,396,995; 6,544,732; 6,401,267; and 7,070,927; WO publications WO 2006/073504 and 2005/082098; and US Pub Nos. 2007/0207482 and 2007/0087362.

[0035] "Nucleic acid", "oligonucleotide", "polynucleotide", "oligo" or grammatical equivalents used herein refers generally to at least two nucleotides covalently linked together. A nucleic acid generally will contain phosphodiester bonds, although in some cases nucleic acid analogs may be included that have alternative backbones such as phosphoramidite, phosphorodithioate, or methylphophoroamidite linkages; or peptide nucleic acid backbones and linkages.

Other analog nucleic acids include those with bicyclic structures including locked nucleic acids, positive backbones, non-ionic backbones and non-ribose backbones. Modifications of the ribose-phosphate backbone may be done to increase the stability of the molecules; for example, PNA: DNA hybrids can exhibit higher stability in some environments.

[0036] "Primer" means an oligonucleotide, either natural or synthetic, that is capable, upon forming a duplex with a polynucleotide template, of acting as a point of initiation of nucleic acid synthesis and being extended from its 3' end along the template so that an extended duplex is formed. The sequence of nucleotides added during the extension process is determined by the sequence of the template polynucleotide. Primers usually are extended by a DNA polymerase.

[0037] "Probe" means generally an oligonucleotide that is complementary to an oligonucleotide or target nucleic acid under investigation. Probes used in certain aspects of the claimed invention are labeled in a way that permits detection, e.g., with a fluorescent or other optically-discernable tag.

[0038] "Sequence determination" in reference to a target nucleic acid means determination of information relating to the sequence of nucleotides in the target nucleic acid. Such information may include the identification or determination of partial as well as full sequence information of the target nucleic acid. The sequence information may be determined with varying degrees of statistical reliability or confidence. In one aspect, the term includes the determination of the identity and ordering of a plurality of contiguous nucleotides in a target nucleic acid starting from different nucleotides in the target nucleic acid.

[0039] "Target nucleic acid" means a nucleic acid from a gene, a regulatory element, genomic DNA, cDNA, RNAs including mRNAs, rRNAs, siRNAs, miRNAs and the like and fragments thereof. A target nucleic acid may be a nucleic acid from a sample, or a secondary nucleic acid such as a product of an amplification reaction.

[0040] As used herein, the term "T_m" is commonly defined as the temperature at which half of the population of double-stranded nucleic acid molecules becomes dissociated into single strands. The equation for calculating the Tm of nucleic acids is well known in the art. As indicated by standard references, a simple estimate of the Tm value may be calculated by the equation: T_m=81.5+16.6(log10[Na+])0.41(% [G+C])−675/n−1.0m, when a nucleic acid is in aqueous solution having cation concentrations of 0.5 M, or less, the (G+C) content is between 30% and 70%, n is the number of bases, and m is the percentage of base pair mismatches (see e.g., Sambrook J et al., "Molecular Cloning, A Laboratory Manual", 3rd Edition, Cold Spring Harbor Laboratory Press (2001)). Other references include more sophisticated computations, which take structural as well as sequence characteristics into account for the calculation of T_m (see also, Anderson and Young (1985), Quantitative Filter Hybridization, *Nucleic Acid Hybridization*, and Allawi and SantaLucia (1997), *Biochemistry* 36:10581-94).

DETAILED DESCRIPTION

[0041] Technology is described herein for providing nucleic acid constructs having interspersed adaptors inserted in a desired orientation with respect to one another for use in large scale sequencing methods. Many adaptor insertion methods developed to date do not allow control of the orientation of newly inserted adaptors vis-à-vis previously inserted adaptors. The inability to control the orientation of adaptors with respect to one another can have a number of undesired consequences. The presence of adaptors in both orientations in a population of target nucleic acid/adaptor library constructs may require multiple sequencing primers in each sequencing reaction to enable sequencing regardless of the orientation of a given adaptor. In addition, analysis of sequence data collected from multiple adaptors of unspecified orientation may require either determination of the orientation of each adaptor or consideration of all possible combinations of adaptor orientation during assembly.

Overview of Sequencing Approaches for use with the Claimed Invention

[0042] FIG. **1** is a simplified flow diagram of an overall method **100** for sequencing nucleic acids using the processes of the claimed invention. Generally, creation of a target molecule for sequencing is accomplished by extracting and preparing target nucleic acids **110** (e.g., fractionating, shearing or cleaving), constructing a library with the sheared target nucleic acids using engineered adaptors **120**, replicating the library constructs to form amplified library constructs (e.g., forming DNA nanoballs through circle dependant replication) **130**, and sequencing the amplified target nucleic acids.

[0043] In process **110** of method **100**, the target nucleic acids for some aspects are derived from genomic DNA. In some aspects such as whole genome sequencing, **10-100** genome-equivalents of DNA preferably are obtained to ensure that the population of target DNA fragments covers the entire genome. The target genomic DNA is isolated using conventional techniques, for example as disclosed in Sambrook and Russell, *Molecular Cloning: A Laboratory Manual*. The target genomic DNA is then fragmented to a desired size by conventional techniques including enzymatic digestion, shearing, or sonication. Fragment size of the target nucleic acid can vary depending on the source target nucleic acid and the library construction methods used, but typically range from 50 nucleotides in length to over 11 kb in length, including 200-700 nucleotides in length, 400-600 nucleotides in length, 450-550 in length, or 4 kb to over 10 kb in length. Alternatively, in some aspects, the target nucleic acids comprise mRNAs or cDNAs. In specific embodiments, the target DNA is created using isolated transcripts from a biological sample. Isolated mRNA may be reverse transcribed into cDNAs using conventional techniques, again as described in *Genome Analysis: A Laboratory Manual Series* (Vols. I-IV) or *Molecular Cloning: A Laboratory Manual.*

[0044] In process **120** of method **100**, a library is constructed using the fragmented target nucleic acids. Library construction will be discussed in detail infra; briefly, the library constructs are assembled by inserting adaptor molecules at a multiplicity of sites throughout each target nucleic acid fragment. The interspersed adaptors permit acquisition of sequence information from multiple sites in the target nucleic acid consecutively or simultaneously. In some aspects, the interspersed adaptors are inserted at intervals within a contiguous region of the target nucleic acids at predetermined positions. The intervals may or may not be equal. In some aspects, the accuracy of the spacing between interspersed adaptors may be known only to an accuracy of one to a few nucleotides. In other aspects, the spacing of the adaptors is known, and the orientation of each adaptor relative to other adaptors in the library constructs is known.

[0045] In process **130** of method **100**, the library constructs are amplified and, in some aspects, are replicated to form DNA nanoballs. In such a process, the library constructs (the target nucleic acids with the interspersed adaptors) are replicated in such a way so as to form single-stranded DNA concatemers of each library construct, each concatamer comprising multiple linear tandem repeats of the library construct. Single-stranded DNA concatemers under conventional conditions (in buffers, e.g., TE, SSC, SSPE or the like) form random coils in a manner known in the art (e.g., see Edvinsson (2002), "On the size and shape of polymers and polymer complexes," Dissertation 696 (University of Uppsala)). Concatemeric DNA randomly coiled forms nanoballs (also termed "DNA nanoballs", "nucleic acid nanoballs" or "DNBs").

[0046] In process **140** of method **100**, the DNBs formed in process **130** are sequenced. In some aspects, the DNBs are randomly arrayed on a planar surface. The DNBs may be covalently or noncovalently attached to the planar surface. The target nucleic acids within each DNB are then sequenced by iterative interrogation using sequencing-by-synthesis techniques and/or sequencing-by-ligation techniques.

[0047] FIG. **2** is a schematic representation of one aspect of a method for assembling adaptor/target nucleic acid library constructs. DNA, such as genomic DNA **202**, is isolated and fragmented **203** into target nucleic acids **204** using standard techniques as described briefly above. The fragmented target nucleic acids **204** are then repaired so that the 5' and 3' ends of each strand are flush or blunt ended. Following this reaction, each fragment is "A-tailed" with a single A added to the 3' end of each strand of the fragmented target nucleic acids using a non-proofreading polymerase **205**. Also as part of process **205**, a first and second arm of a first adaptor is then ligated to each target nucleic acid, producing a target nucleic acid with adaptor arms ligated to each end **206**. In one aspect, the adaptor arms are "T tailed" to be complementary to the A tailing of the target nucleic acid, facilitating ligation of the adaptor arms in a known orientation.

[0048] In a preferred embodiment, the invention provide adaptor ligation to each fragment in a manner that minimizes the creation of intra- or intermolecular ligation artefacts. This is desirable because random fragments of target nucleic acids forming ligation artefacts with one another create false proximal genomic relationships between target nucleic acid fragments, complicating the sequence alignment process. The aspect shown in FIG. **2** shows step **205** as a combination of blunt end repair and an A tail addition. This preferred aspect using both A tailing and T tailing to attach the adaptor to the DNA fragments prevents random intra- or inter-molecular associations of adaptors and fragments, which reduces artefacts that would be created from self-ligation, adaptor-adaptor or fragment-fragment ligation.

[0049] As an alternative to A tailing, various other methods can be implemented to prevent formation of ligation artefacts of the target nucleic acids and the adaptors, as well as orient the adaptor arms with respect to the target nucleic acids, including using complementary NN overhangs in the target nucleic acids and the adaptor arms, or employing blunt end ligation with an appropriate target nucleic acid to adaptor ratio to optimize single fragment nucleic acid/adaptor arm ligation ratios.

[0050] In process **207**, the linear target nucleic acid **206** is circularized, a process that will be discussed in detail infra, resulting in a circular library construct **208** comprising target nucleic acid and an adaptor. Note that the circularization process results in bringing the first and second arms of the first adaptor together to form a contiguous adaptor sequence in the circular construct. In process **209**, the circular construct is amplified, such as by circle dependant amplification, using, e.g., random hexamers and ϕ29 or helicase. Alternatively, target nucleic acid/adaptor structure **206** may remain linear, and amplification may be accomplished by PCR primed from sites in the adaptor arms. The amplification **209** preferably is a controlled amplification process and uses a high fidelity, proof-reading polymerase, resulting in a sequence-accurate library of amplified target nucleic acid/adaptor constructs where there is sufficient representation of the genome or one or more portions of the genome being queried.

[0051] In aspects herein, the first adaptor comprises two Type IIs restriction endonuclease recognition sites, positioned such that the target nucleic acid outside the recognition sequence (and outside of the adaptor) is cut **210**. The arrows around structure **210** indicate the recognition sites and the site of restriction. In process **211**, EcoP15, a Type IIs restriction endonuclease, is used to cut the library constructs. Note that in the aspect shown in FIG. **2**, a portion of each library construct mapping to a portion of the target nucleic acid will be cut away from the construct (the portion of the target nucleic acid between the arrow heads in structure **210**). Restriction of the library constructs with EcoP15 in process **211** results in a library of linear constructs containing the first adaptor, with the first adaptor "interior" to the ends of the linear construct **212**. The resulting linear library construct will have a size defined by the distance between the endonuclease recognition sites and the endonuclease restriction site plus the size of the adaptor. In process **213**, the linear construct **212**, like the fragmented target nucleic acid **204**, is treated by conventional methods to become blunt or flush ended, A tails comprising a single A are added to the 3' ends of the linear library construct using a non-proofreading polymerase and first and second arms of a second adaptor are ligated to ends of the linearized library construct by A-T tailing and ligation **213**. The resulting library construct comprises the structure seen at **214**, with the first adaptor interior to the ends of the linear construct, with target nucleic acid flanked on one end by the first adaptor, and on the other end by either the first or second arm of the second adaptor.

[0052] In process **215**, the double-stranded linear library constructs are treated so as to become single-stranded **216**, and the single-stranded library constructs **216** are then ligated **217** to form single-stranded circles of target nucleic acid interspersed with two adaptors **218**. The ligation/circularization process of **217** is performed under conditions that optimize intramolecular ligation.

[0053] Next, in the two-adaptor aspect shown in FIG. **2**, the single-stranded, circularized library constructs **218** are amplified by circle dependant replication **219** to form DNA nanoballs **220**. Circle dependant replication is performed, e.g., using specific primers where the amplification product displaces its own tail, producing linear, tandem single-stranded copies of |-target nucleic acid/adaptor 1/target nucleic acid/adaptor 2-| library constructs. As the tandem copies begin to multiply, the library constructs begin to coil and form secondary structures, ultimately forming DNA nanoballs. Each library construct contains in some aspects between about ten to about 5000 copies, or from about 250 copies to about 2500 copies of the |-target nucleic acid/adaptor 1/target nucleic acid/adaptor 2-| repeats, and preferably

contains about 500 to about 1200 copies of the the |-target nucleic acid/adaptor 1/target nucleic acid/adaptor 2-| repeats The resulting DNA nanoballs **220**, then, are clonal populations of DNA in discrete structures, which can then be arrayed and sequenced (process not shown).

[0054] FIG. **3** is a simplified schematic illustration showing the cyclical nature of the basic adaptor insertion process **300** where two, three, four, five or more adaptors can be inserted into a target nucleic acid. A fragmented target nucleic acid is shown at **302**. Process **303** provides adaptor arm to target nucleic acid ligation (as was described with some detail in the discussion of the aspect shown in FIG. **2**), resulting in a linear target nucleic acid with first and second adaptor arms of a first adaptor ligated onto its ends **304**. The adaptor arms are then ligated to one another in an intramolecular reaction that results in a circularization of the target nucleic acid/adaptor library construct **306**. The library construct is then amplified **307** resulting in a population comprising a plurality of copies of each target nucleic acid/adaptor library construct **308**. These library constructs **308** are then cleaved **309** (for example, by restriction with a Type IIs restriction endonuclease recognizing one or more sites in the adaptor and cutting in the target nucleic acid sequence), and the cycle continues to add second, third, fourth or more adaptors.

[0055] FIG. **4** is a schematic illustration of one aspect of a DNA array **400** employing multi-adaptor nucleic acid library constructs. The multi-adaptor nucleic acid library constructs in the form of DNA nanoballs (DNBs) are seen at **402**. DNBs are arrayed on a planar matrix **404** having discrete sites **406**. The DNBs **402** may be fixed to the discrete sites by a variety of techniques, including covalent attachment and non-covalent attachment. In one embodiment, the surface of the matrix **406** may comprise attached capture oligonucleotides that form complexes, e.g., double-stranded duplexes, with a segment of an adaptor component of the DNB. In other embodiments, capture oligonucleotides may comprise oligonucleotide clamps, or like structures, that form triplexes with adaptor oligonucleotides (see, e.g., U.S. Pat. No. 5,473,060). In another embodiment, the surface of the array matrix **406** may have reactive functionalities that react with complementary functionalities on the DNBs to form a covalent linkage (see, e.g., Beaucage (2001), *Current Medicinal Chemistry* 8:1213-1244). Once the DNBs are arrayed, the adaptors interspersed in the target nucleic acids are used to acquire sequence information of the target nucleic acids. A variety of sequencing methodologies may be used with multi-adaptor nucleic acid library constructs, including but not limited to hybridization methods as disclosed in U.S. Pat. Nos. 6,864,052; 6,309,824; 6,401,267; sequencing-by-synthesis methods as disclosed in U.S. Pat. Nos. 6,210,891; 6,828,100, 6,833,246; 6,911,345; Margulies, et al. (2005), *Nature* 437: 376-380 and Ronaghi, et al. (1996), *Anal. Biochem.* 242:84-89; and ligation-based methods as disclosed in U.S. Pat. No. 6,306,597; and Shendure et al. (2005) *Science* 309:1728-1739, all of which are incorporated by reference in their entirety.

[0056] In one aspect, the DNBs described herein—particularly those with inserted and interspersed adapters—are used in sequencing by combinatorial probe-anchor ligation reaction (cPAL) (see U.S. Ser. No. 11/679,124, filed Feb. 24, 2007). In brief, cPAL comprises cycling of the following steps: First, an anchor is hybridized to a first adaptor in the DNBs (typically immediately at the 5' or 3' end of one of the adaptors). Enzymatic ligation reactions are then performed with the anchor to a fully degenerate probe population of, e.g., 8-mer probes that are labeled, e.g., with fluorescent dyes. Probes may have a length, e.g., about 6-20 bases, or, preferably, about 7-12 bases. At any given cycle, the population of 8-mer probes that is used is structured such that the identity of one or more of its positions is correlated with the identity of the fluorophore attached to that 8-mer probe. For example, when 7-mer sequencing probes are employed, a set of fluorophore-labeled probes for identifying a base immediately adjacent to an interspersed adaptor may have the following structure: 3'-F1-NNNNNNAp, 3'-F2-NNNNNNGp. 3'-F3-NNNNNNCp and 3'-F4-NNNNNNTp (where "p" is a phosphate available for ligation). In yet another example, a set of fluorophore-labeled 7-mer probes for identifying a base three bases into a target nucleic acid from an interspersed adaptor may have the following structure: 3'-F1-NNNNANNp, 3'-F2-NNNNGNNp. 3'-F3-NNNNCNNp and 3'-F4-NNNNTNNp. To the extent that the ligase discriminates for complementarity at that queried position, the fluorescent signal provides the identity of that base.

[0057] After performing the ligation and four-color imaging, the anchor:8-mer probe complexes are stripped and a new cycle is begun. With T4 DNA ligase, accurate sequence information can be obtained as far as six bases or more from the ligation junction, allowing access to at least 12 bp per adaptor (six bases from both the 5' and 3' ends), for a total of 48 bp per 4-adaptor DNB, 60 bp per 5-adaptor DNB and so on.

[0058] FIG. **5** is a schematic illustration of the components that may be used in an exemplary sequencing-by-ligation technique. A construct **500** is shown with a stretch of target nucleic acid to be analyzed interspersed with three adaptors, with the 5' end of the stretch shown at **502** and the 3' end shown at **504**. The target nucleic acid portions are shown at **506** and **508**, with adaptor **1** shown at **501**, adaptor **2** shown at **503** and adaptor **3** shown at **505**. Four anchors are shown: anchor A**1** (**510**), which binds to the 3' end of adaptor **1** (**501**) and is used to sequence the 5' end of target nucleic acid **506**; anchor A**2** (**512**), which binds to the 5' end of adaptor **2** (**503**) and is used to sequence the 3' end of target nucleic acid **506**; anchor A**3** (**514**), which binds to the 3' end of adaptor **2** (**503**) and is used to sequence the 5' end of target nucleic acid **508**; and anchor A**4** (**516**), which binds to the 5' end of adaptor **3** (**505**) and is used to sequence the 3' end of target nucleic acid **508**.

[0059] Depending on which position that a given cycle is aiming to interrogate, the 8-mer probes are structured differently. Specifically, a single position within each 8-mer probe is correlated with the identity of the fluorophore with which it is labeled. Additionally, the fluorophore molecule is attached to the opposite end of the 8-mer probe relative to the end targeted to the ligation junction. For example, in the graphic shown here, the anchor **530** is hybridized such that its 3' end is adjacent to the target nucleic acid. To query a position five bases into the target nucleic acid, a population of degenerate 8-mer probes shown here at **518** may be used. The query position is shown at **532**. In this case, this correlates with the fifth nucleic acid from the 5' end of the 8-mer probe, which is the end of the 8-mer probe that will ligate to the anchor. In the aspect shown in FIG. **5**, the 8-mer probes are individually labeled with one of four fluorophores, where Cy5 is correlated with A (**522**), Cy3 is correlated with G (**524**), Texas Red is correlated with C (**526**), and FITC is correlated with T (**528**).

[0060] Many different variations of cPAL or other sequencing-by-ligation approaches may be selected depending on various factors such as the volume of sequencing desired, the type of labels employed, the number of different adaptors used within each library construct, the number of bases being queried per cycle, how the DNBs are attached to the surface of the array, the desired speed of sequencing operations, signal detection approaches and the like. In the aspect shown in FIG. **5** and described herein, four fluorophores were used and a single base was queried per cycle. It should, however, be recognized that eight or sixteen fluorophores or more may be used per cycle, increasing the number of bases that can be identified during any one cycle. The degenerate probes (in FIG. **5**, 8-mer probes) can be labeled in a variety of ways, including the direct or indirect attachment of radioactive moieties, fluorescent moieties, colorimetric moieties, chemiluminescent moieties, and the like. Many comprehensive reviews of methodologies for labeling DNA and constructing DNA adaptors provide guidance applicable to constructing oligonucleotide probes of the present invention. Such reviews include Kricka (2002), *Ann. Clin. Biochem.,* 39: 114-129; and Haugland (2006), *Handbook of Fluorescent Probes and Research Chemicals,* 10th Ed. (Invitrogen/Molecular Probes, Inc., Eugene); Keller and Manak (1993), *DNA Probes,* 2nd Ed. (Stockton Press, New York, 1993); and Eckstein (1991), Ed., *Oligonucleotides and Analogues: A Practical Approach* (IRL Press, Oxford); and the like.

[0061] In one aspect, one or more fluorescent dyes are used as labels for the oligonucleotide probes. Labeling can also be carried out with quantum dots, as disclosed in the following patents and patent publications, incorporated herein by reference: U.S. Pat. Nos. 6,322,901; 6,576,291; 6,423,551; 6,251,303; 6,319,426; 6,426,513; 6,444,143; 5,990,479; 6,207,392; 2002/0045045; 2003/0017264; and the like. Commercially available fluorescent nucleotide analogues readily incorporated into the degenerate probes include, for example, Cascade Blue, Cascade Yellow, Dansyl, lissamine rhodamine B, Marina Blue, Oregon Green 488, Oregon Green 514, Pacific Blue, rhodamine 6G, rhodamine green, rhodamine red, tetramethylrhodamine, Texas Red, the Cy fluorophores, the Alexa Fluor® fluorophores, the BODIPY® fluorophores and the like. FRET tandem fluorophores may also be used. Other suitable labels for detection oligonucleotides may include fluorescein (FAM), digoxigenin, dinitrophenol (DNP), dansyl, biotin, bromodeoxyuridine (BrdU), hexahistidine (6×His), phosphor-amino acids (e.g. P-tyr, P-ser, P-thr) or any other suitable label.

[0062] Imaging acquisition may be performed by methods known in the art, such as use of the commercial imaging package Metamorph. Data extraction may be performed by a series of binaries written in, e.g., C/C++, and base-calling and read-mapping may be performed by a series of Matlab and Perl scripts. As described above, for each base in a target nucleic acid to be queried (for example, for 12 bases, reading 6 bases in from both the 5' and 3' ends of each target nucleic acid portion of each DNB), a hybridization reaction, a ligation reaction, imaging and a primer stripping reaction is performed. To determine the identity of each DNB in an array at a given position, after performing the biological sequencing reactions, each field of view ("frame") is imaged with four different wavelengths corresponding to the four fluorescent, e.g., 8-mers used. All images from each cycle are saved in a cycle directory, where the number of images is 4× the number of frames (for example, if a four-fluorophore technique is employed). Cycle image data may then be saved into a directory structure organized for downstream processing.

[0063] Data extraction typically requires two types of image data: bright field images to demarcate the positions of all DNBs in the array; and sets of fluorescence images acquired during each sequencing cycle. The data extraction software identifies all objects with the brightfield images, then for each such object, computes an average fluorescence value for each sequencing cycle. For any given cycle, there are four data-points, corresponding to the four images taken at different wavelengths to query whether that base is an A, G, C or T. These raw base-calls are consolidated, yielding a discontinuous sequencing read for each DNB. The next task is to match these sequencing reads against a reference genome.

[0064] Information regarding the reference genome may be stored in a reference table. A reference table may be compiled using existing sequencing data on the organism of choice. For example human genome data can be accessed through the National Center for Biotechnology Information at ftp.ncbi.nih.gov/refseq/release, or through the J. Craig Venter Institute at http://www.jcvi,org/researchhuref/. All or a subset of human genome information can be used to create a reference table for particular sequencing queries. In addition, specific reference tables can be constructed from empirical data derived from specific populations, including genetic sequence from humans with specific ethnicities, geographic heritage, religious or culturally-defined populations, as the variation within the human genome may slant the reference data depending upon the origin of the information contained therein.

[0065] In an alternative aspect of the claimed invention, parallel sequencing of the target nucleic acids in the DNBs on a random array is performed by combinatorial sequencing-by-hybridization (cSBH), as disclosed by Drmanac in U.S. Pat. Nos. 6,864,052; 6,309,824; and 6,401,267. In one aspect, first and second sets of oligonucleotide probes are provided, where each set has member probes that comprise oligonucleotides having every possible sequence for the defined length of probes in the set. For example, if a set contains probes of length six, then it contains 4096 (4^6) probes. In another aspect, first and second sets of oligonucleotide probes comprise probes having selected nucleotide sequences designed to detect selected sets of target polynucleotides. Sequences are determined by hybridizing one probe or pool of probes, hybridizing a second probe or a second pool or probes, ligating probes that form perfectly matched duplexes on their target nucleic acids, identifying those probes that are ligated to obtain sequence information about the target nucleic acid sequence, repeating the steps until all the probes or pools of probes have been hybridized, and determining the nucleotide sequence of the target nucleic acid from the sequence information accumulated during the hybridization and identification processes.

[0066] In yet another alternative aspect, parallel sequencing of the target nucleic acids in the DNBs is performed by sequencing-by-synthesis techniques as described in U.S. Pat. Nos. 6,210,891; 6,828,100, 6,833,246; 6,911,345; Margulies, et al. (2005), *Nature* 437:376-380 and Ronaghi, et al. (1996), *Anal. Biochem.* 242:84-89. Briefly, modified pyrosequencing, in which nucleotide incorporation is detected by the release of an inorganic pyrophosphate and the generation of

photons, is performed on the DNBs in the array using sequences in the adaptors for binding of the primers that are extended in the synthesis.

Adaptor Insertion and Structure

[0067] The inability to control the orientation of adaptors with respect to one another can have a number of undesired consequences. The presence of adaptors in both orientations in a population of target nucleic acid/adaptor library constructs requires the use of two different anchor oligos in each sequencing reaction to enable sequencing regardless of the orientation of a given adaptor. In addition, sequencing of adaptors of unspecified orientation requires either determination of the orientation of each adaptor—adding at least one additional round of hybridization and scanning to the sequencing process—or consideration of all possible combinations of adaptor orientation during assembly of sequencing reads from adaptors in the same target nucleic acid/adaptor construct.

[0068] FIG. **6** is a schematic illustration of an insertion of a second adaptor relative to a first adaptor in a nucleic acid library construct. Again Process **600** begins with circular library construct **602**, having an inserted first adaptor **610**. First adaptor **610** has a specific orientation, with a rectangle identifying the "outer strand" of the first adaptor and a diamond identifying the "inner strand" of the first adaptor (Ad**1** orientation **610**). A Type IIs restriction endonuclease site in the first adaptor **610** is indicated by the tail of arrow **601**, and the site of cutting is indicated by the arrow head. Process **603** comprises cutting with the Type IIs restriction endonuclease, ligating first and second adaptor arms of a second adaptor, and recircularization. As can be seen in the resulting library constructs **604** and **606**, the second adaptor can be inserted in two different ways relative to the first adaptor. In the desired orientation **604**, the oval is inserted into the circle's outer strand with the rectangle, and the bowtie is inserted into the circle's inner strand with the diamond (Ad**2** orientation **620**). In the undesired orientation the oval is inserted into the circle's inner strand with the diamond and the bowtie is inserted into the circle's outer strand with the rectangle (Ad**2** orientation **630**).

[0069] FIG. **7** is a schematic representation of components of an exemplary adaptor useful for selecting insertion orientation. A basic schematic of an adaptor is shown at **700**. The adaptor comprises a 5' arm **701**, a double-stranded region **702** and a 3' arm **703**. Both the 5' and the 3' arms have a "T tail" **704** and a Type IIs restriction endonuclease site **705** (here, EcoP15). The binding region **702** is the region where the two arms of the adaptor come together to be ligated in the circularization process (**305** of FIG. **3**). Structure **710** is the 5' arm of adaptor **700**. Again, T tail **704** and the EcoP15 site **705** are shown, as well as the 5' arm **701** and the binding region **712**. Structure **720** is the 3' arm of adaptor **700**. Note the T tail **704** and the EcoP15 site **705**, as well as the 3' arm **703** and the binding region **722**. In the 5' arm, the binding region **712** is complementary to the binding region **722** of the 3' arm.

[0070] Because the aspects of the claimed invention work optimally when library constructs are of a desired size and limited target nucleic acid sequence, it is preferred that throughout the library construction process the circularization reactions occur intramolecularly. That is, that the separate constructs of the library that are generated in the library construct assembly cycle (as shown in FIG. **3**) do not ligate to one another. Also, it is preferred that only one set of adaptor arms for each adaptor used in the library construction process be included per target nucleic acid/adaptor construct. Thus, blocking oligos **717** and **727** are used to block the binding regions **712** and **722** regions, respectively. Blocker oligonucleotide **717** is complementary to binding sequence **716**, and blocker olidonucleotide **727** is complementary to binding sequence **726**. In the schematic illustrations of the 5' adaptor arm and the 3' adaptor arm, the underlined bases are ddC and the bolded font bases are phosphorylated. Blocker oligonucleotides **717** and **727** are not covalently bound to the adaptor arms, and can be "melted off" after ligation of the adaptor arms to the library construct and before circularization; further, the dideoxy nucleotide (here, ddC or alternatively a different non-ligatable nucleotide) prevents ligation of blocker to adaptor. In addition or as an alternative, in some aspects, the blocker oligo-adaptor arm hybrids contain a one or more base gap between the adaptor arm and the blocker to reduce ligation of blocker to adaptor. In some aspects, the blocker/binding region hybrids have T_ms of about 37° C. to enable easy melting of the blocker sequences prior to tail to tail ligation (circularization).

[0071] Adaptor structure **730** is a schematic of the final adaptor, where N is an unspecified base, a numeral "1" specifies bases added to disrupt the palindrome (i.e., the EcoP15 site is flanked by A's to isolate the 6-base palindrome formed by the EcoP15 sites on the two arms of the adaptor), numeral "2" specifies bases that correspond to the ddC in the blocker oligonucleotides, numeral "3" specifies the EcoP15 site (CTGCTG) and numeral "4" specifies the T bases designated for TA ligation to the A tailed target nucleic acid. The adaptor shown as **900** and detailed at **930** would, in some aspects, be appropriate for a first adaptor to be added in the construction of a library. Adaptors added subsequently would, in some aspects, have a single Type IIs restriction endonuclease site rather than two sites, and, in some aspects, the Type IIs restriction endonuclease sites in each adaptor would be different from one another. Exemplary Type IIs restriction endonucleases include, but are not limited to, Eco57M I, Mme I, Acu I, Bpm I, BceA I, Bbv I, BciV I, BpuE I, BseM II, BseR I, Bsg I, BsmF I, BtgZ I, Eci I, EcoP15 I, Eco57M I, Fok I, Hga I, Hph I, Mbo II, Mnl I, SfaN I, TspDT I, TspDW I, Taq II, and the like.

[0072] In some aspects, the adaptors when assembled have a total length of about 50 nucleotides. As shown above, in some aspects, the adaptors are ligated to the target nucleic acid as two adaptor arms, where each adaptor arm comprises two adaptor oligos (the two complementary strands) and one blocker oligo. As shown the 5' ends of all four adaptor arm oligos are phosphorylated to support ligation to the insert and tail-to-tail ligation of 5' to 3' adaptor arms. As shown, the 5' and 3' adaptor arms have 3' overhangs at the adaptor-target nucleic acid ligation junctions, to enable ligation to an A-tailed insert, and to suppress head-to-head adaptor arm ligation. Also as shown, the 5' and 3' adaptor arms have Type IIs restriction endonuclease recognition sites oriented to enable cleavage of the adjacent target nucleic acid.

[0073] Again, the adaptor construct shown in FIG. **7** would be, in some aspects, appropriate for a first adaptor to be inserted into a library construct because it contains two Type IIs restriction endonuclease recognition sites. Subsequently inserted adaptors would, in some aspects, comprise a single Type IIs restriction endonuclease recognition site oriented to enable cleavage of the adjacent target nucleic acid. Additionally, in preferred aspects, the 5' and 3' adaptor arms have

anchor primer binding sites to enable sequencing of adjacent target nucleic acids. The anchor primer binding sites in some aspects overlap with the respective Type IIs restriction endonuclease recognition site(s); however, in other aspects the anchor primer binding sites do not overlap with the Type IIs restriction endonuclease recognition site(s).

[0074] FIG. **8** is a schematic representation of adaptor insertion allowing subsequent circularization of the target/adaptor construct. The portion of the library construct seen in FIG. **8** is adaptor-centric, showing target nucleic acid at **802** and **812**, a 5' adaptor arm at **804**, a 5' adaptor arm blocking oligo at **806**, a 3' adaptor arm at **810**, and a 3' adaptor arm blocking oligo at **808**. The T tail of the adaptor arms **804** and **810** and the A tail of the target nucleic acids **802** and **812** are indicated. In process **801**, the adaptor arms are ligated to the target nucleic acid resulting in target nucleic acid/5' adaptor arm structure **814**, and target nucleic acid/3' adaptor arm structure **816**, with blocking oligos **806** and **808** still hybridized to the target nucleic acid/adaptor arm structures. In process **803**, the blockers are removed by melting, and, in preferred aspects under dilute conditions to favor intramolecular ligation of process **805**. The resulting structure is seen at **818**. FIG. **8** illustrates the process of adaptor arm ligation featuring blocking oligos; however, other methods may be used to block ligation-creating concatemers of adaptor arms or of library constructs, including using adaptor arms that comprise a restriction site, preferably a site for a restriction endonuclease that cuts asymmetrically, such as Ava I. Alternatively, the adaptor arms may comprise one or more uracil bases that can be selectively cleaved using uracil-DNA glycosylase enzyme (Krokan et al, 1997) with the resulting fragments then being melted off in the same way the blocker oligo is melted off.

Selection of Adaptor Orientation by Selective Capture

[0075] The claimed methods provide selection of orientation by selective capture (or "hybrid capture"), where a series of purification steps is performed that result in the sequential elimination of various undesired structures. A schematic of this process is shown in FIG. **9**. FIG. **9** shows the results of adding first and second adaptor arms to a library construct that already has a first adaptor interspersed within the target nucleic acid. The first adaptor has a rectangle in the "top strand" and a diamond in the "bottom strand." The second adaptor has an oval in the "top strand" and a bowtie in the "bottom strand." The first arm of the second adaptor is phosphorylated on the 5' end of the top strand (designated by a "P"), whereas the second arm of the second adaptor has a biotin (or other capturable functional moiety) on the 5' end of the bottom strand (designated by a "B"). The functional moiety may be part of the first or second arm of the second adaptor that are ligated to the library construct; alternatively, after ligation of the first and second arms of the second adaptor, the second library constructs may be amplified with a functionalized primer complementary to an end of one strand of the second adaptor, and the functionalized amplified double-stranded second library constructs are captured.

[0076] Ligation of the two arms of the second adaptor results in four combinations: AB (with "P" strand **902** as a top strand and "B" strand **904** as a bottom strand); AA (with "P" strands **902** as both a top strand and a bottom strand; BB (with "B" strands **904** as both a top strand and a bottom strand); and BA (with "B" strand **904** as a top strand and "P" strand **902** as a bottom strand). Selective capture of the biotinylated dsDNA (or otherwise functionalized dsDNA) eliminates unbiotinylated AA structures in **906**. Subsequent denaturation and elution of unbiotinylated ssDNA from the captured dsDNA eliminates double-biotinylated BB structures in **906**. Next, first adaptor top-strand containing ssDNAs (AB) are purified using a first adaptor top-strand specific capture probe (a probe with a sequence specific for a nucleotide sequence of the first-adaptor top strand). The first adaptor top-strand ssDNA structure is then circularized using, e.g., circLigase, and the resulting circles can be amplified and converted to dsDNA using circle dependant amplification.

EXAMPLES

[0077] A Tailing: Samples of 100 ng of fragmented genomic DNA were prepared in Thermopol buffer, with dATP and Taq polymerase added. The samples were then incubated at 70° C. for 60 minutes and cooled to 4° C. The samples were then purified by Qiagen MinElute columns.

[0078] Adaptor annealing: The A tailed fragmented genomic DNA samples were mixed with T tailed adaptors and blocking oligos in a buffer containing NaCl, Tris and EDTA. The samples were then heated to 95° C. for 5 minutes and then allowed to cool to room temperature.

[0079] Adaptor ligation: The annealed adaptor/genomic DNA samples were mixed with HB ligation buffer and T4 ligase. The samples were then incubated at 14° C. for two hours, 70° C. for 10 minutes (to inactivate the T4 enzyme and remove the blocking oligos) and cooled to 4° C. The samples were then purified by Qiagen MinElute columns.

[0080] Adaptor circularization: The linear fragmented genomic DNAs now flanked by first and second arms of an adaptor were circularized by incubation in epicenter buffer and T4 Ligase at 14° C. for 14 hours. The samples were then heat inactivated at 70° C. for 10 minutes and then cooled to 4° C.

[0081] The present specification provides a complete description of the methodologies, systems and/or structures and uses thereof in example aspects of the presently-described technology. Although various aspects of this technology have been described above with a certain degree of particularity, or with reference to one or more individual aspects, those skilled in the art could make numerous alterations to the disclosed aspects without departing from the spirit or scope of the technology hereof. Since many aspects can be made without departing from the spirit and scope of the presently described technology, the appropriate scope resides in the claims hereinafter appended. Other aspects are therefore contemplated. Furthermore, it should be understood that any operations may be performed in any order, unless explicitly claimed otherwise or a specific order is inherently necessitated by the claim language. It is intended that all matter contained in the above description and shown in the accompanying drawings shall be interpreted as illustrative only of particular aspects and are not limiting to the embodiments shown. Changes in detail or structure may be made without departing from the basic elements of the present technology as defined in the following claims.

SEQUENCE LISTING

<160> NUMBER OF SEQ ID NOS: 9

<210> SEQ ID NO 1
<211> LENGTH: 15
<212> TYPE: DNA
<213> ORGANISM: Artificial Sequence
<220> FEATURE:
<223> OTHER INFORMATION: Synthetic oligonucleotide

<400> SEQUENCE: 1

acucuagcug acuag 15

<210> SEQ ID NO 2
<211> LENGTH: 35
<212> TYPE: DNA
<213> ORGANISM: Artificial Sequence
<220> FEATURE:
<223> OTHER INFORMATION: Target nucleic acid
<220> FEATURE:
<221> NAME/KEY: misc_feature
<222> LOCATION: (5)...(21)
<223> OTHER INFORMATION: N is A, C, G or T

<400> SEQUENCE: 2

gagtnnnnnn nnnnnnnnnn tgagatcgac tgatc 35

<210> SEQ ID NO 3
<211> LENGTH: 30
<212> TYPE: DNA
<213> ORGANISM: Artificial Sequence
<220> FEATURE:
<223> OTHER INFORMATION: Synthetic oligonucleotide

<400> SEQUENCE: 3

actgctgacg cttacgatgc acgatacgtc 30

<210> SEQ ID NO 4
<211> LENGTH: 32
<212> TYPE: DNA
<213> ORGANISM: Artificial Sequence
<220> FEATURE:
<223> OTHER INFORMATION: Synthetic oligonucleotide

<400> SEQUENCE: 4

ttgacgactg cgaatgctac gtgctatgca gt 32

<210> SEQ ID NO 5
<211> LENGTH: 32
<212> TYPE: DNA
<213> ORGANISM: Artificial Sequence
<220> FEATURE:
<223> OTHER INFORMATION: Synthetic oligonucleotide

<400> SEQUENCE: 5

tgcacgatac gtctacgatg cgaacagcag at 32

<210> SEQ ID NO 6
<211> LENGTH: 30
<212> TYPE: DNA
<213> ORGANISM: Artificial Sequence
<220> FEATURE:
<223> OTHER INFORMATION: Synthetic oligonucleotide

-continued

<400> SEQUENCE: 6

cgtgctatgc agatgctacg cttgtcgtct 30

<210> SEQ ID NO 7
<211> LENGTH: 50
<212> TYPE: DNA
<213> ORGANISM: Artificial Sequence
<220> FEATURE:
<223> OTHER INFORMATION: Synthetic oligonucleotide
<220> FEATURE:
<221> NAME/KEY: misc_feature
<222> LOCATION: (10)...(19)
<223> OTHER INFORMATION: N is A, G, C or T
<220> FEATURE:
<221> NAME/KEY: misc_feature
<222> LOCATION: (21)...(30)
<223> OTHER INFORMATION: N is A, G, C or T
<220> FEATURE:
<221> NAME/KEY: misc_feature
<222> LOCATION: (32)...(41)
<223> OTHER INFORMATION: N is A, G, C or T

<400> SEQUENCE: 7

aactgctgan nnnnnnnnng nnnnnnnnnn cnnnnnnnnn nacagcagat 50

<210> SEQ ID NO 8
<211> LENGTH: 49
<212> TYPE: DNA
<213> ORGANISM: Artificial Sequence
<220> FEATURE:
<223> OTHER INFORMATION: Synthetic oligonucleotide

<400> SEQUENCE: 8

aactgctgac gcttacgatg cacgatacgt ctacgatgcg aacagcaga 49

<210> SEQ ID NO 9
<211> LENGTH: 49
<212> TYPE: DNA
<213> ORGANISM: Artificial Sequence
<220> FEATURE:
<223> OTHER INFORMATION: Synthetic oligonucleotide

<400> SEQUENCE: 9

tgacgactgc gaatgctacg tgctatgcag atgctacgct tgtcgtcta 49

What is claimed is:

1. A method for selecting for orientation of two adaptors with respect to one another in library constructs comprising:

(a) obtaining target nucleic acid;

(b) ligating a first adaptor to the target nucleic acid to produce first library constructs, wherein one strand of the first adaptor comprises a capture sequence;

(c) ligating first and second arms of a second adaptor to the linearized first library constructs to form second library constructs, wherein at least one strand of one of the second adaptor arms comprises a functional group;

(d) capturing functionalized double-stranded second library constructs and discarding un-functionalized second library constructs;

(e) eluting single-stranded nucleic acids without a functional group from the captured double-stranded functionalized second library constructs; and

(f) capturing the capture sequence in the one strand of the first adaptor, thereby selecting for library constructs having a desired orientation of the second adaptor with respect to the first adaptor.

2. The method of claim **1**, further comprising repeating processes (b) through (f) until a desired number of adaptors have been inserted into the nucleic acid library constructs.

3. The method of claim **1**, wherein the first library constructs are cut with a restriction endonuclease after being circularized.

4. The method of claim **1**, wherein the first adaptor is ligated to the target nucleic acid as two adaptor arms.

5. The method of claim **1**, wherein the functional group is biotin and the functionalized double-stranded second library constructs are captured by a strepavidin column.

6. The method of claim **1**, wherein the first and second adaptors further comprise Type IIs endonuclease recognition sites.

7. A method for selecting for orientation of two adaptors with respect to one another in library constructs comprising:
(a) obtaining a target nucleic acid;
(b) ligating a first adaptor to the target nucleic acid to produce first library constructs, wherein one strand of the first adaptor comprises a capture sequence;
(c) ligating first and second arms of a second adaptor to the linearized first library constructs to form second library constructs;
(d) amplifying the second library constructs with a functionalized primer complementary to an end of one strand of the second adaptor;
(d) capturing functionalized amplified double-stranded second library constructs and discarding unfunctionalized second library constructs;
(e) eluting single-stranded nucleic acids without a functional group from the captured double-stranded functionalized second library constructs; and
(f) capturing the capture sequence in the one strand of the first adaptor, thereby selecting for library constructs having a desired orientation of the second adaptor with respect to the first adaptor.

8. The method of claim 7, wherein the first library constructs are cut with a restriction endonuclease after being circularized.

9. The method of claim 7, wherein each adaptor is ligated to the target nucleic acid as two adaptor arms.

10. The method of claim 7, wherein the functional group is biotin and the functionalized double-stranded second library constructs are captured by a strepavidin column.

11. The method of claim 7, wherein each adaptor further comprises one or more endonuclease recognition sites.

12. The method of claim 11, wherein the endonuclease recognition sites are Type IIs endonuclease recognition sites.

13. An amplicon made by amplification of a circular library construct comprising target nucleic acid interspersed with a plurality of adaptors, wherein at least one of the plurality of adaptors has a desired orientation with respect to at least one of the other of the plurality of adaptors.

14. The amplicon of claim 13, wherein each of the plurality of adaptors has a desired orientation with respect to at least one other of the plurality of adaptors.

15. The amplicon of claim 13, wherein one or more of the adaptors comprises a restriction endonuclease recognition site.

16. The amplicon of claim 15, wherein the restriction endonuclease recognition site is a Type IIs restriction endonuclease recognition site.

17. The amplicon of claim 13, wherein each adaptor of the plurality of adaptors further comprise a different anchor primer binding site at a 5' and 3' end of each of the plurality of adaptors.

18. A multiplicity of amplicons of circular library constructs, wherein each amplicon comprises target nucleic acid interspersed with a plurality of adaptors, wherein at least one of the plurality of adaptors has a desired orientation with respect to at least one of the other of the plurality of adaptors.

19. The multiplicity of amplicons of claim 18, wherein each of the plurality of adaptors has a desired orientation with respect to at least one other of the plurality of adaptors.

20. The multiplicity of amplicons of claim 18, wherein the target nucleic acid is genomic DNA, cDNA or RNA, and wherein the multiplicity of amplicons comprises substantially all of genomic DNA, cDNA or RNA of interest.

21. The multiplicity of amplicons of claim 18, wherein one or more of the adaptors comprises a restriction endonuclease recognition site.

22. The multiplicity of amplicons of claim 21, wherein the restriction endonuclease recognition site is a Type IIs restriction endonuclease recognition site.

23. The multiplicity of amplicons of claim 18, wherein each adaptor of the plurality of adaptors further comprise a different anchor primer binding site at a 5' and 3' end of each of the plurality of adaptors.

24. A kit for inserting interspersed adaptors in target nucleic acid, wherein said kit comprises:
b) a first double-stranded adaptor;
c) a functionalized second double stranded adaptor;
d) reagents for capturing functionalized second double-stranded adaptor; and
(e) reagents for capturing one strand of the first double-stranded adaptor.

25. The kit of claim 24 further comprising:
a) a ligase;
b) a first Type IIs restriction endonuclease;
c) a second Type IIs restriction endonuclease; or
d) a functionalized third adaptor.

* * * * *