Attorney Docket No. 067784-5026WO

EFFICIENT BASE DETERMINATION IN SEQUENCING REACTIONS

CROSS REFERENCE TO RELATED APPLICATIONS

[0001] This application claims the benefit of priority to U.S. Patent Application Nos: 60/992,485, filed December 5, 2007; 61/026,337, filed February 5, 2008; 61/035,914, filed March 12, 2008; 61/061,134, filed June 13, 2008; 61/116,193, filed November 19, 2008; 61/102,586, filed on October 3, 2008; 12/265,593, filed November 5, 2008; and 12/266,385, filed November 6, 2008, each of which is hereby incorporated by reference in its entirety.

BACKGROUND OF THE INVENTION

[0002] Large-scale genomic sequence analysis is a key step toward understanding a wide range of biological phenomena. 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 multiple nucleic acid targets simultaneously.

[0003] Conventional methods of sequencing are generally restricted to determining a few tens of nucleotides before signals become significantly degraded, thus placing a significant limit on overall sequencing efficiency. Conventional methods of sequencing are also often limited by signal-to-noise ratios that render such methods unsuitable for single-molecule sequencing.

[0004] It would be advantageous for the field if methods and compositions could be designed to increase the efficiency of sequencing reactions as well as the efficiency of assembling complete sequences from shorter read lengths.

SUMMARY OF THE INVENTION

[0005] Accordingly, the present invention provides methods and compositions for sequencing reactions.

[0006] In one aspect, the present invention provides a method for determining a sequence in a target nucleic acid. This method includes the steps of: (a) providing a sequencing template that includes a fragment of the target nucleic acid and an adaptor that includes at least a first anchor site; (b) hybridizing an anchor probe to the anchor site, and the anchor probe includes a region complementary to the adaptor site and three or more degenerate bases for binding in the target nucleic acid sequence; (c) hybridizing a pool of sequencing probes for determination of the sequence of one or more nucleotides in a defined position relative to the adaptor, wherein the sequencing probe is detectably labeled to identify the presence of a particular base; (d) ligating the anchor probe and the sequencing probe; and (e) detecting the sequencing probe, thereby determining a sequence in the target nucleic acid.

[0007] In a further aspect, the invention provides a method of determining the identification of a first nucleotide at a detection position of a target sequence comprising a plurality of detection positions. This method includes the steps of: (a) providing a surface with a plurality of concatemers, wherein each concatemer includes a plurality of monomers and each monomer comprises: (i) a first target domain of the target sequence comprising a first set of target detection positions; (ii) at least a first adaptor

comprising: (1) a first anchor site; and (2) a second adjacent anchor site; (b) hybridizing a first anchor probe to the first anchor site; (c) hybridizing a second anchor probe to the second anchor site, wherein the second anchor probe also hybridizes to sequences outside the second anchor site; (d) hybridizing at least a first sequencing probe to the first target domain, wherein the first sequencing probe comprises: (i) a first probe domain complementary to the target domain; (ii) a unique nucleotide at a first interrogation position; and (iii) a label; under conditions wherein if the unique nucleotide is complementary to the first nucleotide, the sequencing probe hybridizes to the concatemer; (e) ligating the anchor probes and the sequencing probe; and (f) identifying the first nucleotide.

BRIEF DESCRIPTION OF THE DRAWINGS

[0008] Figure 1 is a schematic illustration of an embodiment of a method for fragmenting nucleic acids.

[0009] Figure 2 is a schematic illustration of embodiments of the invention related to long fragment read (LFR) technology. Figure 2A illustrates a method for fragmenting nucleic acids by a standard multiple displacement amplification method (MDA). Figure 2B illustrates a method for fragmenting nucleic acids by a multiple displacement amplification method utilizing a 5' exonuclease. Figure 2C is a schematic of an embodiment of the overall LFR process.

[0010] Figure 3 is a schematic illustration of embodiments of a barcode adaptor design for use in methods of the invention.

[0011] Figure 4 is a schematic illustration of an embodiment of the invention for fragmenting nucleic acids utilizing a nick translation method.

[0012] Figure 5 is a schematic illustration of adaptors that can be used in embodiments of the invention. Figure 5A provides four different sequences of adaptors. Figure 5B illustrates different components that can be included in the design of an adaptor of the invention.

[0013] Figure 6 is a schematic illustration of an embodiment of the invention for making circular nucleic acid templates comprising multiple adaptors.

[0014] Figure 7 is a schematic illustration of an embodiment of the invention for controlling the orientation of adaptors inserted into target nucleic acids.

[0015] Figure 8 is a schematic illustration of exemplary embodiments of different orientations in which adaptors and target nucleic acid molecules can be ligated to each other.

[0016] Figure 9 is a schematic illustration of one aspect of a method for assembling nucleic acid templates of the invention.

[0017] Figure 10 is a schematic illustration of components of adaptors that are useful for controlling the way such adaptors are inserted into a target nucleic acid.

[0018] Figure 11 is a schematic illustration of an embodiment of an arm-by-arm ligation process for inserting adaptors into target nucleic acids. Figure 11A illustrates an exemplary embodiment of the arm-by-arm ligation process and Figure 11B illustrates exemplary components of adaptor arms of use in this process.

[0019] Figure 12 is a schematic illustration of possible orientations of adaptor insertion.

[0020] Figure 13 is a schematic illustration of one embodiment of a nick translation ligation method.

[0021] Figure 14 is a schematic illustration of one embodiment of a method for inserting multiple adaptors.

[0022] Figure 15 is a schematic illustration of one embodiment of a nick translation ligation method.

[0023] Figure 16 is a schematic illustration of one embodiment of a nick translation ligation method.

[0024] Figure 17 is a schematic illustration of one embodiment of a nick translation ligation method utilizing nick translation circle inversion (Figure 17A) and nick translation circle inversion combined with uracil degradation (Figure 17B).

[0025] Figure 18 is a schematic illustration of an embodiment of a nick translation ligation method.

[0026] Figure 19 is a schematic illustration of one embodiment of a method for inserting multiple adaptors.

[0027] Figure 20 is a schematic illustration of one embodiment of a method for inserting multiple adaptors.

[0028] Figure 21 is a schematic illustration of one embodiment of a method for inserting multiple adaptors.

[0029] Figure 22 is a schematic illustration of one embodiment of a method for inserting multiple adaptors.

[0030] Figure 23 is a schematic illustration of one embodiment of a combinatorial probe anchor ligation method.

[0031] Figure 24 is a schematic illustration of one embodiment of a combinatorial probe anchor ligation method.

[0032] Figure 25 is a schematic illustration of one embodiment of a combinatorial probe anchor ligation method.

[0033] Figure 26 is a schematic illustration of one embodiment of a combinatorial probe anchor ligation method.

[0034] Figure 27 is a graph of fluorescent intensity levels achieved for each base at defined positions using a double combinatorial probe anchor ligation method.

[0035] Figure 28 is a graph of data fit scores obtained for interrogated positions using a combinatorial probe anchor ligation method.

[0036] Figure 29 is a graph of fluorescent intensity levels obtained for a single base interrogation at different time points using a single and a double combinatorial probe anchor ligation method.

[0037] Figure 30 is a graph of data fit scores obtained for a single base interrogation at different time points using a single combinatorial probe anchor ligation method.

[0038] Figure 31 is a graph of fluorescent intensity levels achieved for different positions using a variety of second anchor probes in double combinatorial probe anchor ligation methods compared to a single combinatorial probe anchor ligation method.

[0039] Figure 32 is a graph illustrating the data fit scores obtained for different positions using a variety of second anchor probes in double combinatorial probe anchor ligation methods compared to a single combinatorial probe anchor ligation method.

[0040] Figure 33 is a graph illustrating the fluorescent intensity levels achieved for different positions using a variety of double combinatorial probe anchor ligation methods compared to a single combinatorial probe anchor ligation method.

[0041] Figure 34 is a graph illustrating data fit scores obtained for different positions using a variety of first anchor probes of different length in a double combinatorial probe anchor ligation method.

[0042] Figure 35 is a graph illustrating the fluorescent intensity levels achieved for each base at a defined position using a double combinatorial probe anchor ligation method in the presence of a kinase at different temperatures.

[0043] Figure 36 is a graph illustrating data fit scores obtained for a defined position using a double combinatorial probe anchor ligation method in the presence of a kinase at different temperatures.

[0044] Figure 37 is a graph illustrating the fluorescent intensity levels achieved for each base at a defined position using a double combinatorial probe anchor ligation method in the presence of a kinase at different kinase incubation times.

[0045] Figure 38 is a graph illustrating the data fit scores obtained for a defined position using a double combinatorial probe anchor ligation method in the presence of a kinase at different kinase incubation times.

DETAILED DESCRIPTION OF THE INVENTION

[0046] The practice of the present invention may employ, unless otherwise indicated, conventional techniques and descriptions of organic chemistry, polymer technology, molecular biology (including recombinant techniques), cell biology, biochemistry, and immunology, which are within the skill of the art. Such conventional techniques include polymer array synthesis, hybridization, ligation, and detection of hybridization using a label. Specific illustrations of suitable techniques can be had by reference to the example herein below. 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 *Genome Analysis: A Laboratory Manual Series* (Vols. I-IV), *Using Antibodies: A Laboratory Manual, Cells: A Laboratory Manual, PCR Primer: A Laboratory Manual, and Molecular Cloning: A Laboratory Manual* (all from Cold Spring Harbor Laboratory Press), Stryer, L. (1995) *Biochemistry* (4th Ed.) Freeman, New York, 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.

[0047] 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 "a polymerase" refers to one agent or mixtures of such agents, and reference to "the method" includes reference to equivalent steps and methods known to those skilled in the art, and so forth.

[0048] 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, compositions, formulations and methodologies which are described in the publication and which might be used in connection with the presently described invention.

[0049] Where a range of values is provided, it is understood that each intervening value, to the tenth of the unit of the lower limit unless the context clearly dictates otherwise, 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 is 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.

[0050] 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.

[0051] Although the present invention is described primarily with reference to specific embodiments, it is also envisioned that other embodiments will become apparent to those skilled in the art upon reading the present disclosure, and it is intended that such embodiments be contained within the present inventive methods.

I. Overview

[0052] The present invention is directed to compositions and methods for nucleic acid identification and detection, which find use in a wide variety of applications as described herein.

[0053] The overall method for sequencing target nucleic acids using the compositions and methods of the present invention includes extracting and fragmenting target nucleic acids from a sample. The fragmented nucleic acids are used to produce target nucleic acid templates that will generally include one or more adaptors. The target nucleic acid templates are subjected to amplification methods to form nucleic acid nanoballs, which are usually disposed on a surface. Sequencing applications are performed on the nucleic acid nanoballs of the invention, usually through sequencing by ligation techniques, including combinatorial probe anchor ligation ("cPAL") methods, which are described in further detail below. cPAL and other sequencing methods can also be used to detect specific sequences, such as including single nucleotide polymorphisms ("SNPs") in nucleic acid constructs of the invention, (which include nucleic acid nanoballs as well as linear and circular nucleic acid templates).

II. Preparing fragments of genomic nucleic acid

[0054] As discussed further herein, nucleic acid templates of the invention comprise target nucleic acids and adaptors. In order to obtain target nucleic acids for construction of the nucleic acid templates of the invention, the present invention provides methods for obtaining genomic nucleic acids from a sample and for fragmenting those genomic nucleic acids to produce fragments of use in subsequent methods for constructing nucleic acid templates of the invention.

IIA. Overview of preparing fragments of genomic nucleic acid

[0055] Target nucleic acids can be obtained from a sample using methods known in the art. As will be appreciated, the sample may comprise any number of substances, including, but not limited to, bodily fluids (including, but not limited to, blood, urine, serum, lymph, saliva, anal and vaginal secretions, perspiration and semen, of virtually any organism, with mammalian samples being preferred and human samples being particularly preferred); environmental samples (including, but not limited to, air, agricultural, water and soil samples); biological warfare agent samples; research samples (i.e. in the case of nucleic acids, the sample may be the products of an amplification reaction, including both target

and signal amplification as is generally described in PCT/US99/01705, such as PCR amplification reaction); purified samples, such as purified genomic DNA, RNA, proteins, etc.; raw samples (bacteria, virus, genomic DNA, etc.); as will be appreciated by those in the art, virtually any experimental manipulation may have been done on the sample. In one aspect, the nucleic acid constructs of the invention are formed from genomic DNA. In certain embodiments, the genomic DNA is obtained from whole blood or cell preparations from blood or cell cultures.

[0056] In an exemplary embodiment, genomic DNA is isolated from a target organism. By "target organism" is meant an organism of interest and as will be appreciated, this term encompasses any organism from which nucleic acids can be obtained, particularly from mammals, including humans, although in some embodiments, the target organism is a pathogen (for example for the detection of bacterial or viral infections). Methods of obtaining nucleic acids from target organisms are well known in the art. Samples comprising genomic DNA of humans find use in many embodiments. In some aspects such as whole genome sequencing, about 20 to about 1,000,0000 or more genome-equivalents of DNA are preferably obtained to ensure that the population of target DNA fragments sufficiently covers the entire genome. The number of genome equivalents obtained may depend in part on the methods used to further prepare fragments of the genomic DNA for use in accordance with the present invention. For example, in the long fragment read methods described further below, about 20 to about 50 genome equivalents are generally utilized. For methods utilizing multiple displacement amplification, which is also described further below, about 1000 to about 100,000 genome equivalents are generally utilized. For methods in which no amplification is used prior to fragmenting, about 100,000 to about 1,000,000 genome equivalents are used.

[0057] The target genomic DNA is isolated using conventional techniques, for example as disclosed in Sambrook and Russell, *Molecular Cloning: A Laboratory Manual,* cited *supra*. The target genomic DNA is then fractionated or fragmented to a desired size by conventional techniques including enzymatic digestion, shearing, or sonication, with the latter two finding particular use in the present invention.

[0058] Fragment sizes 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 to 600 nucleotides in length. In another embodiment, the fragments are 300 to 600 or 200 to 2000 nucleotides in length. In yet another embodiment, the fragments are 10-100, 50-100, 50-300, 100-200, 200-300, 50-400, 100-400, 200-400, 300-400, 400-500, 400-600, 500-600, 50-1000, 100-1000, 200-1000, 300-1000, 400-1000, 500-1000, 600-1000, 700-1000, 700-900, 700-800, 800-1000, 900-1000, 1500-2000, 1750-2000, and 50-2000 nucleotides in length.

[0059] In a further embodiment, fragments of a particular size or in a particular range of sizes are isolated. Such methods are well known in the art. For example, gel fractionation can be used to produce a population of fragments of a particular size within a range of basepairs, for example for 500 base pairs $\pm$ 50 base pairs.

[0060] In many cases, enzymatic digestion of extracted DNA is not required because shear forces created during lysis and extraction will generate fragments in the desired range. In a further embodiment, shorter fragments (1-5 kb) can be generated by enzymatic fragmentation using restriction endonucleases. In a still further embodiment, about 10 to about 1,000,000 genome-equivalents of DNA ensure that the population of fragments covers the entire genome. Libraries containing nucleic acid

templates generated from such a population of fragments will thus comprise target nucleic acids whose sequences, once identified and assembled, will provide most or all of the sequence of an entire genome.

[0061] In some cases, it is advantageous to provide carrier DNA, e.g. unrelated circular synthetic double- stranded DNA, to be mixed and used with the sample DNA whenever only small amounts of sample DNA are available and there is danger of losses through nonspecific binding, e.g. to container walls and the like.

[0062] In one embodiment, the DNA is denatured after fragmentation to produce single stranded fragments.

[0063] In one embodiment, after fragmenting, (and in fact before or after any step outlined herein) an amplification step can be applied to the population of fragmented nucleic acids to ensure that a large enough concentration of all the fragments is available for subsequent steps of creating the decorated nucleic acids of the invention and using those nucleic acids for obtaining sequence information. Such amplification methods are well known in the art and include without limitation: polymerase chain reaction (PCR), ligation chain reaction (sometimes referred to as oligonucleotide ligase amplification OLA), cycling probe technology (CPT), strand displacement assay (SDA), transcription mediated amplification (TMA), nucleic acid sequence based amplification (NASBA), rolling circle amplification (RCA) (for circularized fragments), and invasive cleavage technology.

[0064] In further embodiments, after fragmenting, target nucleic acids are further modified to prepare them for insertion of multiple adaptors according to methods of the invention. Such modifications can be necessary because the process of fragmentation may result in target nucleic acids with termini that are not amenable to the procedures used to insert adaptors, particularly the use of enzymes such as ligases and polymerases. As for all the steps outlined herein, this step is optional and can be combined with any step.

[0065] In an exemplary embodiment, after physical fragmenting, target nucleic acids frequently have a combination of blunt and overhang ends as well as combinations of phosphate and hydroxyl chemistries at the termini. In this embodiment, the target nucleic acids are treated with several enzymes to create blunt ends with particular chemistries. In one embodiment, a polymerase and dNTPs is used to fill in any 5' single strands of an overhang to create a blunt end. Polymerase with 3' exonuclease activity (generally but not always the same enzyme as the 5' active one, such as T4 polymerase) is used to remove 3' overhangs. Suitable polymerases include, but are not limited to, T4 polymerase, Taq polymerases, E. coli DNA Polymerase 1, Klenow fragment, reverse transcriptases, Φ29 related polymerases including wild type Φ29 polymerase and derivatives of such polymerases, T7 DNA Polymerase, T5 DNA Polymerase, RNA polymerases. These techniques can be used to generate blunt ends, which are useful in a variety of applications.

[0066] In further optional embodiments, the chemistry at the termini is altered to avoid target nucleic acids from ligating to each other. For example, in addition to a polymerase, a protein kinase can also be used in the process of creating blunt ends by utilizing its 3' phosphatase activity to convert 3' phosphate groups to hydroxyl groups. Such kinases can include without limitation commercially available kinases such as T4 kinase, as well as kinases that are not commercially available but have the desired activity.

[0067] Similarly, a phosphatase can be used to convert terminal phosphate groups to hydroxyl groups. Suitable phosphatases include, but are not limited to, Alkaline Phosphatase (including Calf Intestinal (CIP)), Antarctic Phosphatase, Apyrase, Pyrophosphatase, Inorganic (yeast) thermostable inorganic

pyrophosphatase, and the like, which are known in the art and commercially available, for example from New England Biolabs.

[0100] As depicted in Figure 16, these modifications prevent the target nucleic acids from ligating to each other in later steps of methods of the invention, thus ensuring that during steps in which adaptors (and/or adaptor arms) are ligated to the termini of target nucleic acids, target nucleic acids will ligate to adaptors but not to other target nucleic acids. Target nucleic acids 1601 and 1602 are preferably ligated to adaptors 1603 and 1604 in a desired orientation (as illustrated in this figure, the desired orientation is one in which the ends with the same shape –circle or square – ligates to each other). Modifying the ends avoids the undesired configurations 1607, 1608, 1609 and 1610, in which the target nucleic acids ligate to each other and the adaptors ligate to each other. In addition, as will be discussed in further detail below, the orientation of each adaptor-target nucleic acid ligation can also be controlled through control of the chemistry of the termini of both the adaptors and the target nucleic acids. The control over the chemistry of the termini can be provided using methods known in the art and described further herein.

[0101] As will be appreciated by those in the art, and as for all the steps outlined herein, any combination of these steps and enzymes may be used. For example, some enzymatic fragmentation techniques, such as the use of restriction endonucleases, may render one or more of these enzymatic "end repair" steps superfluous.

[0102] The modifications described above can prevent the creation of nucleic acid templates containing different fragments ligated in an unknown conformation, thus reducing and/or removing the errors in sequence identification and assembly that can result from such undesired templates.

[0068] In some cases, hierarchical fragmentation methods are utilized in combination with any of the enzymatic or mechanical methods of fragmenting described herein. Such methods are described in US Application No. 11/451,692 and published PCT application WO 2006/138284, which are herein incorporated by reference in their entirety for all purposes and in particular for all teachings related to hierarchical fragmentation.

[0069] In some embodiments, a controlled random enzymatic ("CoRE") fragmentation method is utilized to prepare fragments for use in the invention. CoRE fragmentation is an enzymatic endpoint assay, and has the advantages of enzymatic fragmentation (such as the ability to use it on low amounts and/or volumes of DNA) without many of its drawbacks (including sensitivity to variation in substrate or enzyme concentration and sensitivity to digestion time). Briefly, CoRE fragmentation involves a series of three enzymatic steps, which are schematically illustrate in Figure 1. First, a nucleic acid 101 is subjected to an enzyme catalyzed multiple displacement amplification (MDA) in the presence of dNTPs doped with dUTP or UTP in a defined ration to the dTTP. This results in the substitution of deoxyuracil ("dU") or uracil ("U") at defined and controllable proportions of the T positions in both strands of the amplification product (103). The U moieties are then excised (104), usually through a combination of UDG, EndoVIII, and T4PNK, to create single base gaps with functional 5' phosphate and 3' hydroxyl ends (105). The single base gaps will be created at an average spacing defined by the frequency of U in the MDA product. Treatment of the gapped nucleic acid (105) with a polymerase results in nick translation until nicks on opposite strands converge, thereby creating double strand breaks, resulting a relatively population of double stranded fragments of a relatively homogenous size (107). Since the size distribution of the double stranded fragments (107) is a result of the ration of dTTP to DUTP or UTP used

in the MDA reaction, rather than by the duration or degree of enzymatic treatment, this CoRE fragmentation methods produces high degrees of fragmentation reproducibility.

[0070] In some cases, particularly when it is desired to isolate long fragments (such as fragments from about 150 to about 750 kilobases in length), the present invention provides methods in which cells are lysed and the intact nucleic are pelleted with a gentle centrifugation step. The genomic nucleic acid, usually genomic DNA, is released through enzymatic digestion, using for example proteinase K and RNase digestion over several hours. The resultant material is then dialyzed overnight or diluted directly to lower the concentration of remaining cellular waste. Since such methods of isolating the nucleic acid does not involve many disruptive processes (such as ethanol precipitation, centrifugation, and vortexing), the genomic nucleic acid remains largely intact, yielding a majority of fragments in excess of 150 kilobases.

[0071] In some cases, in combination with any of the above-described fragmentation methods, the present invention further provides methods of aliquoting a population of genomic nucleic acid fragments which allows the reconstruction of diploid genomes, e.g. the identification of maternal and paternal chromosomes or sequence. This is a significant advantage over the processes of the prior art.

[0072] In this embodiment, the genomic fragments are aliqoted such that the nucleic acids are diluted to a concentration of approximately 10% of a haploid genome per aliquot. At such a level of dilution, approximately 95% of the base pairs in a particular aliquot are non-overlapping. This method of aliquoting, also referred to herein as a long fragment read (LFR) fragmentation method, can in particular embodiments be used on large molecular weight fragments isolated according to the methods described above and further herein. An example of an LFR method is schematically illustrated in Figure 2C. LFR usually begins with a short treatment of genomic nucleic acids, usually genomic DNA, with a 5' exonuclease to create 3' single-stranded overhangs. Such single stranded overhangs serve as multiple displacement amplification (MDA) initiation sites (Figure 2A). The 5' exonuclease treated DNA is then diluted to sub-genome concentrations and dispersed across a number of aliquots, usually across a number of wells in a multiwell plate. The fragments in each well are amplified, usually using a standard MDA method (Figure 2A) and/or an MDA method utilizing an exonuclease (Figure 2B). In some cases, the amplification method introduces uracil moieties into the fragments, such that the above-described CoRE method can be used to further fragment the fragments in each well following amplification. The fragmenting of the MDA products can also be achieved via sonication or enzymatic treatment. In general, following fragmentation of the MDA products, the ends of the resultant fragments are repaired, usually with T4 polymerase and T4 polynucleotide kinase. Fragments are then treated with alkaline phosphatase and then the fragments are tagged with an adaptor. Generally, the tag adaptor arm is designed in two segments – one segment is common to all wells and blunt end ligates directly to the fragments using methods described further herein. The second segment is unique to each well and contains a "barcode" sequence such that when the contents of each well are combined, the fragments from each well can be identified. Figure 3 illustrates some exemplary barcode adaptors that can be added to the fragments for this aspect of the invention.

[0073] In some cases, an LFR method is used to analyze the genome of an individual cell. The process for isolating DNA in this case is similar to the methods described above, but occurs in a smaller volume. Once the DNA is isolated and before it is aliquoted into individual wells, the genomic DNA must be carefully fragmented to avoid loss of material, particularly to avoid loss of sequence from the ends of

each fragment, since loss of such material will result in gaps in the final genome assembly. In some cases, sequence loss is avoided through use of an infrequent nicking enzyme, which creates starting sites for a polymerase, such as phi29 polymerase, at distances of approximately 100 kb from each other. As the polymerase creates the new DNA strand, it displaces the old strand, with the end result being that there are overlapping sequences near the sites of polymerase initiation (Figure 4), resulting in very few deletions of sequence.

[0074] In some cases, it is advantageous to provide carrier DNA, e.g. unrelated circular synthetic double- stranded DNA, to be mixed and used with the sample DNA whenever only small amounts of sample DNA are available and there is danger of losses through nonspecific binding, e.g. to container walls and the like. In one embodiment, the DNA is denatured after fragmentation to produce single stranded fragments.

[0075] In one embodiment, after fragmenting, (and in fact before or after any step outlined herein) an amplification step can be applied to the population of fragmented nucleic acids to ensure that a large enough concentration of all the fragments is available for subsequent steps of creating the decorated nucleic acids of the invention and using those nucleic acids for obtaining sequence information. Such amplification methods are well known in the art and include without limitation: polymerase chain reaction (PCR), ligation chain reaction (sometimes referred to as oligonucleotide ligase amplification OLA), cycling probe technology (CPT), strand displacement assay (SDA), transcription mediated amplification (TMA), nucleic acid sequence based amplification (NASBA), rolling circle amplification (RCA) (for circularized fragments), and invasive cleavage technology.

[0076] In further embodiments, after fragmenting, target nucleic acids are further modified to prepare them for insertion of multiple adaptors according to methods of the invention. Such modifications can be necessary because the process of fragmentation may result in target nucleic acids with termini that are not amenable to the procedures used to insert adaptors, particularly the use of enzymes such as ligases and polymerases. As for all the steps outlined herein, this step is optional and can be combined with any step. Methods for modifying the fragments to prepare them for directed ligation to other nucleic acid molecules include application of enzymes, such as polymerases and phosphatases, to modify the ends of the fragments such that they are only able to ligate to other nucleic acid molecules in a desired orientation. Such methods are described further herein.

IIB. CoRE fragmentation

[0077] As discussed above, methods of fragmentation for use in the present invention include both mechanical and enzymatic fragmentation methods, as well as combinations of enzymatic and fragmentation methods. Many mechanical and enzymatic fragmentation methods are well known in the art.

[0078] In one aspect, the present invention provides a method of fragmentation referred to herein as Controlled Random Enzymatic (CoRE) fragmentation. The CoRE fragmentation methods described herein can be used alone or in combination with other mechanical and enzymatic fragmentation methods known in the art. CoRE fragmentation involves a series of three enzymatic steps, which are schematically illustrated in Figure 1. First, a nucleic acid 101 is subjected to an amplification method that is conducted in the present of dNTPs doped with a proportion of deoxyuracil (“dU”) or uracil (“U”) to result

in substitution of dUTP or UTP at defined and controllable proportions of the T positions in both strands of the amplification product (103). A number of amplification methods can be used in this step of the invention, including without limitation polymerase chain reaction (PCR), ligation chain reaction (sometimes referred to as oligonucleotide ligase amplification OLA), cycling probe technology (CPT), strand displacement assay (SDA), transcription mediated amplification (TMA), nucleic acid sequence based amplification (NASBA), rolling circle amplification (RCA) (for circularized fragments), and invasive cleavage technology. In certain embodiment, multiple displacement amplification (MDA) in the presence of dNTPs doped with dUTP or UTP in a defined ratio to the dTTP is used to create amplification products with dUTP or UTP substituted into certain points on both strands (103).

[0079] After amplification and insertion of the uracil moieties, the uracils are then excised (104), usually through a combination of UDG, EndoVIII, and T4PNK, to create single base gaps with functional 5' phosphate and 3' hydroxyl ends (105). The single base gaps will be created at an average spacing defined by the frequency of U in the MDA product. That is, the higher the amount of dUTP, the shorter the resulting fragments. As will be appreciated by those in the art, other techniques that will result in selective replacement of a nucleotide with a modified nucleotide that can similarly result in cleavage can also be used, such as chemically or other enzymatically susceptible nucleotides.

[0080] Treatment of the gapped nucleic acid (105) with a polymerase with exonuclease activity results in "translation" or "translocation" of the nicks along the length of the nucleic acid until nicks on opposite strands converge, thereby creating double strand breaks, resulting a relatively population of double stranded fragments of a relatively homogenous size (107). The exonuclease activity of the polymerase (such as Taq polymerase) will excise the short DNA strand that abuts the nick while the polymerase activity will "fill in" the nick and subsequent nucleotides in that strand (essentially, the Taq moves along the strand, excising bases using the exonuclease activity and adding the same bases, with the result being that the nick is translocated along the strand until the enzyme reaches the end).

[0081] Since the size distribution of the double stranded fragments (107) is a result of the ration of dTTP to DUTP or UTP used in the MDA reaction, rather than by the duration or degree of enzymatic treatment, this CoRE fragmentation methods produces high degrees of fragmentation reproducibility. Thus, CoRE fragmentation results in a population of double stranded nucleic acid fragments that are all of a similar size.

IIC. Long Fragment Read technology

[0082] Long Fragment Read (LFR) methods of the invention are based on the physical separation of long genomic DNA fragments across many different aliquots such that the probability of any given region of the genome of both the maternal and paternal component in the same aliquot is very rare. By placing a unique identifier in each aliquot and analyzing many aliquot in the aggregate, long fragments of DNA can be assembled into a diploid genome, e.g. the sequence of each parental chromosome can be provided, providing a significant advantage over the prior art. While the discussion herein focuses on the use of the LFR methods using DNB arrays and sequencing by ligation, it should be appreciated that these LFR methods can be used with a variety of other arrays and other sequencing methods to result in the sequencing of diploid genomes as two separate haploid genomes. This can facilitate the identification of familial genetic disease, etc.

[0083] To achieve an appropriate separation of fragments, in general the DNA is diluted to a concentration of approximately 10% of a haploid genome per aliquot (Figure 2C). At this concentration, 95% of the base pairs in an aliquot are non-overlapping. Such a dilution results in a statistical separation such that maternal and paternal fragments will usually land in different aliquots (Figure 2C, second panel). It should be appreciated that the dilution factor can depend on the original size of the fragments. That is, using gentle techniques to isolate genomic DNA, fragments of roughly 100kb can be obtained, which are then aliquoted. Techniques that allow larger fragments result in a need for fewer aliquots, and those that result in shorter fragments may require more dilution.

[0084] In some embodiments, the fragments in each aliquot are amplified and in further embodiments the fragments in each aliquot are further fragmented and then tagged with an adaptor such that fragments from the same aliquot will all comprise the same tag adaptor; see for example US 2007/0072208, hereby incorporated by reference in its entirety, and in particular for the discussions of additional aliquoting and coverage.

[0085] In many embodiments, each aliquot is contained in a separate well of a multi-well plate (for example, a 384 well plate). It will be appreciated that although the following discussion of LFR is provided in terms of a multi-well plate, that any number of different types of containers and systems can be used to hold the different aliquots generated in this method. Such containers and systems are well known in the art and it would be apparent to one of skill in the art what types of containers and systems would be appropriate to use in accordance with this aspect of the invention.

[0086] As discussed above, isolating long fragments of genomic nucleic acid from a cell can be accomplished by a number of different methods. In one embodiment, cells are lysed and the intact nucleic are pelleted with a gentle centrifugation step. The genomic DNA is then released through proteinase K and RNase digestion for several hours. The material can then in some embodiments be treated to lower the concentration of remaining cellular waste – such treatments are well known in the art and can include without limitation dialysis for a period of time (i.e., from 2 -16 hours) and/or dilution. Since such methods of isolating the nucleic acid does not involve many disruptive processes (such as ethanol precipitation, centrifugation, and vortexing), the genomic nucleic acid remains largely intact, yielding a majority of fragments that have lengths in excess of 150 kilobases. In some embodiments, the fragments are from about 100 to about 750 kilobases in lengths. In further embodiments, the fragments are from about 150 to about 600, about 200 to about 500, about 250 to about 400, and about 300 to about 350 kilobases in length.

[0087] An example of an LFR method is schematically illustrated in Figure 2. LFR usually begins with a short treatment of genomic nucleic acids, usually genomic DNA, with a 5' exonuclease to create 3' single-stranded overhangs. Such single stranded overhangs serve as MDA initiation sites (Figure 2). The use of the exonuclease also eliminates the need for a heat or alkaline denaturation step prior to amplification without introducing bias into the population of fragments. In some embodiments, alkaline denaturation is combined with the 5' exonuclease treatment, which results in a reduction in bias that is greater than what is seen with either treatment alone.

[0088] The DNA treated with the 5' exonuclease and optionally the alkaline denaturation is then diluted to sub-genome concentrations and dispersed across a number of aliquots, usually across a number of wells in a multiwell plate. In some embodiments, a 10% genome equivalent is aliquoted into each well of a multiwell plate. If a 384 well plate is used, a 10% genome equivalent aliquot into each well results in

each plate comprising 38 genomes in total. In further embodiments, a 5-50% genome equivalent is aliquoted into each well. As noted above, the number of aliquots and genome equivalents can depend on the original fragment size.

[0089] After separation across multiple wells, the fragments in each well are amplified, usually using an MDA method. In certain embodiments, the MDA reaction is a modified Phi29 polymerase-based amplification reaction. Although much of the discussion herein is in terms of an MDA reaction, it will be appreciated by those of skill in the art that many different kinds of amplification reactions can be used in accordance with the present invention, and that such amplification reactions are well known in the art and described generally in Maniatis et al., Molecular Cloning: A Laboratory Manual, 2d Edition, 1989, and Short Protocols in Molecular Biology, ed. Ausubel, et al, hereby incorporated by reference.

[0090] In some embodiments, the MDA reaction is designed to introduce uracils into the amplification products. In some embodiments, a standard MDA reaction utilizing random hexamers is used to amplify the fragments in each well. In many embodiments, rather than the random hexamers, random 8-mer primers are used to reduce amplification bias in the population of fragments. In further embodiments, several different enzymes can also be added to the MDA reaction to reduce the bias of the amplification. For example, low concentrations of non-processive 5' exonucleases and/or single-stranded binding proteins can be used to create binding sites for the 8-mers. Chemical agents such as betaine, DMSO, and trehalose can also be used to reduce bias through similar mechanisms.

[0091] After amplification of the fragments in each well, the amplification products are then in many embodiments subjected to another round of fragmentation. In some embodiments the above-described CoRE method is used to further fragment the fragments in each well following amplification. As discussed above, in order to use the CoRE method, the MDA reaction used to amplify the fragments in each well is designed to incorporate uracils into the MDA products. The fragmenting of the MDA products can also be achieved via sonication or enzymatic treatment. Enzymatic treatment that could be used in this embodiment includes the use of without limitation DNase I, T7 endonuclease I, Micrococcal nuclease, and the like.

[0092] In some embodiments, when a CoRE method is used to fragment the MDA products, each well containing MDA products are treated with a mix of Uracil DNA glycosylase (UDG), DNA glycosylase-lyase Endonuclease VIII, and T4 polynucleotide kinase to excise the uracil bases and create single base gaps with functional 5' phosphate and 3' hydroxyl groups. Nick translation through use of a polymerase such as Taq polymerase results in double stranded blunt end breaks, resulting in ligatable fragments of a size range dependent on the concentration of dUTP added in the MDA reaction. In some embodiments, the CoRE method used involves removing uracils by polymerization and strand displacement by phi29 (see Figure 4).

[0093] In general, following fragmentation of the MDA products, the ends of the resultant fragments are repaired. Such repairs can be necessary, because many fragmentation techniques can result in termini with overhanging ends and termini with functional groups that are not useful in later ligation reactions, such as 3' and 5' hydroxyl groups and/or 3' and 5' phosphate groups. In many aspects of the present invention, it is useful to have fragments that are repaired to have blunt ends, and in some cases, it can be desirable to alter the chemistry of the termini such that the correct orientation of phosphate and hydroxyl groups is not present, thus preventing "polymerization" of the target sequences. The control over the chemistry of the termini can be provided using methods known in the art. For example, in some

circumstances, the use of phosphatase eliminates all the phosphate groups, such that all ends contain hydroxyl groups. Each end can then be selectively altered to allow ligation between the desired components. One end of the fragments can then be "activated", in some embodiments by treatment with alkaline phosphatase. The fragments are then in many embodiments tagged with an adaptor. In general, these tag adaptors can be used to identify fragments that come from the same well in the LFR method.

[0094] Figure 3 provides a schematic illustration of some embodiments of adaptor design for use as a tag in accordance with the LFR method. Generally, the adaptor is designed in two segments – one segment is common to all wells and blunt end ligates directly to the fragments using methods described further herein. In the embodiment pictured in Figure 3, the "common" adaptor is added as two adaptor arms – one arm is blunt end ligated to the 5' end of the fragment and the other arm is blunt end ligated to the 3' end of the fragment. The second segment of the tagging adaptor is a "barcode" segment that is unique to each well. This barcode is generally a unique sequence of nucleotides, and each fragment in a particular well is given the same barcode. Thus, when the tagged fragments from all the wells are re-combined for sequencing applications, fragments from the same well can be identified through identification of the barcode adaptor. In the embodiment illustrated in Figure 3, the barcode is ligated to the 5' end of the common adaptor arm. The common adaptor and the barcode adaptor can be ligated to the fragment sequentially or simultaneously. As will be described in further detail herein, the ends of the common adaptor and the barcode adaptor can be modified such that each adaptor segment will ligate in the correct orientation and to the proper molecule. Such modifications prevent "polymerization" of the adaptor segments or the fragments by ensuring that the fragments are unable to ligate to each other and that the adaptor segments are only able to ligate in the illustrated orientation.

[0095] In further embodiments, a three segment design is utilized for the adaptors used to tag fragments in each well. This embodiment is similar to the barcode adaptor design described above, except that the barcode adaptor segment is split into two segments (see Figure 3). This design allows for a wider range of possible barcodes by allowing combinatorial barcode adaptor segments to be generated by ligating different barcode segments together to form the full barcode segment. This combinatorial design provides a larger repertoire of possible barcode adaptors while reducing the number of full size barcode adaptors that need to be generated.

[0096] After the fragments in each well are tagged, all of the fragments are combined to form a single population. These fragments can then be used to generate nucleic acid templates of the invention, as is discussed in further detail below. The nucleic acid templates generated from these tagged fragments will be identifiable as belonging to a particular well by the barcode tag adaptors attached to each fragment.

[0097] In some cases, an LFR method is used to analyze the genome of a small number of cells, including an individual cell. The process for isolating DNA in this case is similar to the methods described above, but occurs in a smaller volume. Once the DNA is isolated and before the step of aliquoting the sample into individual wells, the genomic DNA must be carefully fragmented to avoid loss of material, particularly to avoid loss of sequence from the ends of each fragment, since loss of such material will result in gaps in the final genome assembly. In some cases, sequence loss is avoided through use of an infrequent nicking enzyme, which creates starting sites for a polymerase, such as phi29 polymerase, at distances of approximately 100 kb from each other. As the polymerase creates the new DNA strand, it displaces the old strand, with the end result being that there are overlapping

sequences near the sites of polymerase initiation (Figure 4), resulting in very few deletions of sequence. In a further embodiment, the DNA can then be diluted and aliquoted into multiple wells following the methods described above. In some embodiments, a controlled use of a 5' exonuclease (either before or during the MDA reaction) can promote multiple replications of the original DNA from the single cell and thus minimize propagation of early errors through copying of copies.

[0098] It will be appreciated that the LFR methods described herein can be used for sequencing diploid genomes using any sequencing methods known in the art. The LFR methods described herein may in further embodiments be used on any number of sequencing platforms, including for example without limitation: GeneChip (Affymetrix), CodeLink Bioarray (Amersham), Expression Array System (Applied Biosystems), SurePrint microarrays (Agilent), Sentrix LD BeadChip or Sentrix Array Matrix (Illumina) and Verigene (Nanosphere).

[0099] In some embodiments, LFR methods described herein do not include multiple levels or tiers of fragmentation/aliquoting, as described in U.S. Patent Application No. 11/451,692, filed June 13, 2006, which is herein incorporated by reference in its entirety for all purposes and in particular for all teachings related to methods of fragmenting and aliquoting nucleic acids. That is, some embodiments utilize only a single round of aliquoting, and also allow the repooling of aliquots for a single array, rather than using separate arrays for each aliquot.

III. Nucleic acid templates of the invention

[0103] The present invention provides nucleic acid template comprising target nucleic acids and multiple interspersed adaptors. The nucleic acid template constructs are assembled by inserting adaptors molecules at a multiplicity of sites throughout each target nucleic acid. The interspersed adaptors permit acquisition of sequence information from multiple sites in the target nucleic acid consecutively or simultaneously.

[0104] The term "target nucleic acid" refers to a nucleic acid of interest. In one aspect, target nucleic acids of the invention are genomic nucleic acids, although other target nucleic acids can be used, including mRNA (and corresponding cDNAs, etc.). Target nucleic acids include naturally occurring or genetically altered or synthetically prepared nucleic acids (such as genomic DNA from a mammalian disease model). Target nucleic acids can be obtained from virtually any source and can be prepared using methods known in the art. For example, target nucleic acids can be directly isolated without amplification, isolated by amplification using methods known in the art, including without limitation polymerase chain reaction (PCR), strand displacement amplification (SDA), multiple displacement amplification (MDA), rolling circle amplification (RCA), rolling circle amplification (RCR) and other amplification methodologies. Target nucleic acids may also be obtained through cloning, including but not limited to cloning into vehicles such as plasmids, yeast, and bacterial artificial chromosomes.

[0105] In some aspects, the target nucleic acids comprise mRNAs or cDNAs. In certain 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*.

[0106] The target nucleic acids may be single stranded or double stranded, as specified, or contain portions of both double stranded or single stranded sequence. Depending on the application, the nucleic

acids may be DNA (including genomic and cDNA), RNA (including mRNA and rRNA) or a hybrid, where the nucleic acid contains any combination of deoxyribo- and ribo-nucleotides, and any combination of bases, including uracil, adenine, thymine, cytosine, guanine, inosine, xathanine hypoxathanine, isocytosine, isoguanine, etc.

[0107] By “nucleic acid” or "oligonucleotide" or “polynucleotide” or grammatical equivalents herein means at least two nucleotides covalently linked together. A nucleic acid of the present invention will generally contain phosphodiester bonds, although in some cases, as outlined below (for example in the construction of primers and probes such as label probes), nucleic acid analogs are included that may have alternate backbones, comprising, for example, phosphoramide (Beaucage et al., Tetrahedron 49(10):1925 (1993) and references therein; Letsinger, J. Org. Chem. 35:3800 (1970); Sprinzl et al., Eur. J. Biochem. 81:579 (1977); Letsinger et al., Nucl. Acids Res. 14:3487 (1986); Sawai et al, Chem. Lett. 805 (1984), Letsinger et al., J. Am. Chem. Soc. 110:4470 (1988); and Pauwels et al., Chemica Scripta 26:141 91986)), phosphorothioate (Mag et al., Nucleic Acids Res. 19:1437 (1991); and U.S. Pat. No. 5,644,048), phosphorodithioate (Briu et al., J. Am. Chem. Soc. 111:2321 (1989), O-methylphophoroamidite linkages (see Eckstein, Oligonucleotides and Analogues: A Practical Approach, Oxford University Press), and peptide nucleic acid (also referred to herein as “PNA”) backbones and linkages (see Egholm, J. Am. Chem. Soc. 114:1895 (1992); Meier et al., Chem. Int. Ed. Engl. 31:1008 (1992); Nielsen, Nature, 365:566 (1993); Carlsson et al., Nature 380:207 (1996), all of which are incorporated by reference). Other analog nucleic acids include those with bicyclic structures including locked nucleic acids (also referred to herein as “LNA”), Koshkin et al., J. Am. Chem. Soc. 120:13252 3 (1998); positive backbones (Denpcy et al., Proc. Natl. Acad. Sci. USA 92:6097 (1995); non-ionic backbones (U.S. Pat. Nos. 5,386,023, 5,637,684, 5,602,240, 5,216,141 and 4,469,863; Kiedrowshi et al., Angew. Chem. Intl. Ed. English 30:423 (1991); Letsinger et al., J. Am. Chem. Soc. 110:4470 (1988); Letsinger et al., Nucleoside & Nucleotide 13:1597 (1994); Chapters 2 and 3, ASC Symposium Series 580, "Carbohydrate Modifications in Antisense Research", Ed. Y. S. Sanghui and P. Dan Cook; Mesmaeker et al., Bioorganic & Medicinal Chem. Lett. 4:395 (1994); Jeffs et al., J. Biomolecular NMR 34:17 (1994); Tetrahedron Lett. 37:743 (1996)) and non-ribose backbones, including those described in U.S. Pat. Nos. 5,235,033 and 5,034,506, and Chapters 6 and 7, ASC Symposium Series 580, "Carbohydrate Modifications in Antisense Research", Ed. Y. S. Sanghui and P. Dan Cook. Nucleic acids containing one or more carbocyclic sugars are also included within the definition of nucleic acids (see Jenkins et al., Chem. Soc. Rev. (1995) pp 169 176). Several nucleic acid analogs are described in Rawls, C & E News Jun. 2, 1997 page 35. “Locked nucleic acids” (LNA™) are also included within the definition of nucleic acid analogs. LNAs are a class of nucleic acid analogues in which the ribose ring is “locked” by a methylene bridge connecting the 2’-O atom with the 4’-C atom. All of these references are hereby expressly incorporated by reference in their entirety for all purposes and in particular for all teachings related to nucleic acids. These modifications of the ribose-phosphate backbone may be done to increase the stability and half-life of such molecules in physiological environments. For example, PNA:DNA and LNA-DNA hybrids can exhibit higher stability and thus may be used in some embodiments.

[0108] The nucleic acid templates (also referred to herein as “nucleic acid constructs” and “library constructs”) of the invention comprise target nucleic acids and adaptors. As used herein, the term “adaptor” refers to an oligonucleotide of known sequence. Adaptors of use in the present invention may

include a number of elements. The types and numbers of elements (also referred to herein as "features") included in an adaptor will depend on the intended use of the adaptor. Adaptors of use in the present invention will generally include without limitation sites for restriction endonuclease recognition and/or cutting, particularly Type IIs recognition sites that allow for endonuclease binding at a recognition site within the adaptor and cutting outside the adaptor as described below, sites for primer binding (for amplifying the nucleic acid constructs) or anchor primer (sometimes also referred to herein as "anchor probes") binding (for sequencing the target nucleic acids in the nucleic acid constructs), nickase sites, and the like. In some embodiments, adaptors will comprise a single recognition site for a restriction endonuclease, whereas in other embodiments, adaptors will comprise two or more recognition sites for one or more restriction endonucleases. As outlined herein, the recognition sites are frequently (but not exclusively) found at the termini of the adaptors, to allow cleavage of the double stranded constructs at the farthest possible position from the end of the adaptor.

[0109] In some embodiments, adaptors of the invention have a length of about 10 to about 250 nucleotides, depending on the number and size of the features included in the adaptors. In certain embodiments, adaptors of the invention have a length of about 50 nucleotides. In further embodiments, adaptors of use in the present invention have a length of about 20 to about 225, about 30 to about 200, about 40 to about 175, about 50 to about 150, about 60 to about 125, about 70 to about 100, and about 80 to about 90 nucleotides.

[0110] In further embodiments, adaptors may optionally include elements such that they can be ligated to a target nucleic acid as two "arms". One or both of these arms may comprise an intact recognition site for a restriction endonuclease, or both arms may comprise part of a recognition site for a restriction endonuclease. In the latter case, circularization of a construct comprising a target nucleic acid bounded at each termini by an adaptor arm will reconstitute the entire recognition site.

[0111] In still further embodiments, adaptors of use in the invention will comprise different anchor binding sites at their 5' and the 3' ends of the adaptor. As described further herein, such anchor binding sites can be used in sequencing applications, including the combinatorial probe anchor ligation (cPAL) method of sequencing, described herein and in U.S. Application Nos. 60/992,485; 61/026,337; 61/035,914; 61/061,134; 61/116,193; 61/102,586; 12/265,593; and 12/266,385 11/938,106; 11/938,096; 11/982,467; 11/981,804; 11/981,797; 11/981,793; 11/981,767; 11/981,761;11/981,730; 11/981,685; 11/981,661;11/981,607; 11/981,605; 11/927,388; 11/927,356;11/679,124;11/541,225;10/547,214; and 11/451,691, all of which are hereby incorporated by reference in their entirety, and particularly for disclosure relating to sequencing by ligation.

[0112] In one aspect, adaptors of the invention are interspersed adaptors. By "interspersed adaptors" is meant herein oligonucleotides that are inserted at spaced locations within the interior region of a target nucleic acid. In one aspect, "interior" in reference to a target nucleic acid means a site internal to a target nucleic acid prior to processing, such as circularization and cleavage, that may introduce sequence inversions, or like transformations, which disrupt the ordering of nucleotides within a target nucleic acid.

[0113] The nucleic acid template constructs of the invention contain multiple interspersed adaptors inserted into a target nucleic acid, and in a particular orientation. As discussed further herein, the target nucleic acids are produced from nucleic acids isolated from one or more cells, including one to several million cells. These nucleic acids are then fragmented using mechanical or enzymatic methods.

[0114] The target nucleic acid that becomes part of a nucleic acid template construct of the invention may have interspersed adaptors 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. That is, in many embodiments, the adaptors are inserted at known distances, such that the target sequence on one termini is contiguous in the naturally occurring genomic sequence with the target sequence on the other termini. For example, in the case of a Type IIs restriction endonuclease that cuts 16 bases from the recognition site, located 3 bases into the adaptor, the endonuclease cuts 13 bases from the end of the adaptor. Upon the insertion of a second adaptor, the target sequence "upstream" of the adaptor and the target sequence "downstream" of the adaptor are actually contiguous sequences in the original target sequence.

[0115] Although the embodiments of the invention described herein are generally described in terms of circular nucleic acid template constructs, it will be appreciated that nucleic acid template constructs may also be linear. Furthermore, nucleic acid template constructs of the invention may be single- or double-stranded, with the latter being preferred in some embodiments

[0116] The present invention provides nucleic acid templates comprising a target nucleic acid containing one or more interspersed adaptors. In a further embodiment, nucleic acid templates formed from a plurality of genomic fragments can be used to create a library of nucleic acid templates. Such libraries of nucleic acid templates will in some embodiments encompass target nucleic acids that together encompass all or part of an entire genome. That is, by using a sufficient number of starting genomes (e.g. cells), combined with random fragmentation, the resulting target nucleic acids of a particular size that are used to create the circular templates of the invention sufficiently "cover" the genome, although as will be appreciated, on occasion, bias may be introduced inadvertently to prevent the entire genome from being represented.

[0117] The nucleic acid template constructs of the invention comprise multiple interspersed adaptors, and in some aspects, these interspersed adaptors comprise one or more recognition sites for restriction endonucleases. In further aspect, the adaptors comprise recognition sites for Type IIs endonucleases. Type-IIs endonucleases are generally commercially available and are well known in the art. Like their Type-II counterparts, Type-IIs endonucleases recognize specific sequences of nucleotide base pairs within a double stranded polynucleotide sequence. Upon recognizing that sequence, the endonuclease will cleave the polynucleotide sequence, generally leaving an overhang of one strand of the sequence, or "sticky end." Type-IIs endonucleases also generally cleave outside of their recognition sites; the distance may be anywhere from about 2 to 30 nucleotides away from the recognition site depending on the particular endonuclease. Some Type-IIs endonucleases are "exact cutters" that cut a known number of bases away from their recognition sites. In some embodiments, Type IIs endonucleases are used that are not "exact cutters" but rather cut within a particular range (e.g. 6 to 8 nucleotides). Generally, Type IIs restriction endonucleases of use in the present invention have cleavage sites that are separated from their recognition sites by at least six nucleotides (i.e. the number of nucleotides between the end of the recognition site and the closest cleavage point). 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, MnI I, SfaN I, TspDT I, TspDW I,

Taq II, and the like. In some exemplary embodiments, the Type IIs restriction endonucleases used in the present invention are AcuI, which has a cut length of about 16 bases with a 2-base 3' overhang and EcoP15, which has a cut length of about 25 bases with a 2-base 5' overhang. As will be discussed further below, the inclusion of a Type IIs site in the adaptors of the nucleic acid template constructs of the invention provides a tool for inserting multiple adaptors in a target nucleic acid at a defined location.

[0118] As will be appreciated, adaptors may also comprise other elements, including recognition sites for other (non-Type IIs) restriction endonucleases, primer binding sites for amplification as well as binding sites for probes used in sequencing reactions ("anchor probes"), described further herein.

[0119] In one aspect, adaptors of use in the invention have sequences as shown in Figure 5. As identified in the schematic of one of the adaptors in Figure 5B, adaptors can comprise multiple functional features, including recognition sites for Type IIs restriction endonucleases (503 and 506), sites for nicking endonucleases (504) as well as sequences that can influence secondary characteristics, such as bases to disrupt hairpins (501 and 502). Adaptors of use in the invention may in addition contain palindromic sequences, which can serve to promote intramolecular binding once nucleic acid templates comprising such adaptors are used to generate concatemers, as is discussed in more detail below.

IV. Preparing nucleic acid templates of the invention

IVA. Overview of generation of circular templates

[0120] The present invention is directed to compositions and methods for nucleic acid identification and detection, which finds use in a wide variety of applications as described herein, including a variety of sequencing and genotyping applications. The methods described herein allow the construction of circular nucleic acid templates that are used in amplification reactions that utilize such circular templates to create concatamers of the monomeric circular templates, forming "DNA nanoballs", described below, which find use in a variety of sequencing and genotyping applications. The circular or linear constructs of the invention comprise target nucleic acid sequences, generally fragments of genomic DNA (although as described herein, other templates such as cDNA can be used), with interspersed exogeneous nucleic acid adaptors. The present invention provides methods for producing nucleic acid template constructs in which each subsequent adaptor is added at a defined position and also optionally in a defined orientation in relation to one or more previously inserted adaptors. These nucleic acid template constructs are generally circular nucleic acids (although in certain embodiments the constructs can be linear) that include target nucleic acids with multiple interspersed adaptors. These adaptors, as described below, are exogenous sequences used in the sequencing and genotyping applications, and usually contain a restriction endonuclease site, particularly for enzymes such as Type IIs enzymes that cut outside of their recognition site. For ease of analysis, the reactions of the invention preferably utilize embodiments where the adaptors are inserted in particular orientations, rather than randomly. Thus the invention provides methods for making nucleic acid constructs that contain multiple adaptors in particular orientations and with defined spacing between them.

[0121] Control over the spacing and orientation of insertion of each subsequent adaptor provides a number of advantages over random insertion of interspersed adaptors. In particular, the methods described herein improve the efficiency of the adaptor insertion process, thus reducing the need to

introduce amplification steps as each subsequent adaptor is inserted. In addition, controlling the spacing and orientation of each added adaptor ensures that the restriction endonuclease recognition sites that are generally included in each adaptor are positioned to allow subsequent cleavage and ligation steps to occur at the proper point in the nucleic acid construct, thus further increasing efficiency of the process by reducing or eliminating the formation of nucleic acid templates that have adaptors in the improper location or orientation. In addition, control over location and orientation of each subsequently added adaptor can be beneficial to certain uses of the resultant nucleic acid construct, because the adaptors serve a variety of functions in sequencing applications, including serving as a reference point of known sequence to aid in identifying the relative spatial location of bases identified at certain positions within the target nucleic acid. Such uses of adaptors in sequencing applications are described further herein.

[0122] Genomic nucleic acid, generally double stranded DNA (601 in Figure 6), is obtained from a plurality of cells, generally from about 10 to 100 to 1000 to more cells. The use of a plurality of cells allows the ultimate DNA nanoballs to have a level of redundancy which allows for good sequencing coverage of the genome. The genomic nucleic acid is fractionated into appropriate sizes using standard techniques such as physical or enzymatic fractionation combined with size fractionation as is described herein.

[0123] The 5' and 3' ends of the double stranded fragments can optionally be adjusted, as described herein. For example, many techniques used to fractionate nucleic acids result in a combination of lengths and chemistries on the termini of the fragments. For example, the termini may contain overlaps, and for many purposes, blunt ends of the double stranded fragments are preferred. This can be done using known techniques such as a polymerase and dNTPs. Similarly, the fractionation techniques may also result in a variety of termini, such as 3' and 5' hydroxyl groups and/or 3' and 5' phosphate groups. In some embodiments, as described below, it is desirable to enzymatically alter these termini. For example, to prevent the ligation of multiple fragments without the adaptors, it can be desirable to alter the chemistry of the termini such that the correct orientation of phosphate and hydroxyl groups is not present, thus preventing "polymerization" of the target sequences. The control over the chemistry of the termini can be provided using methods known in the art. For example, in some circumstances, the use of phosphatase eliminates all the phosphate groups, such that all ends contain hydroxyl groups. Each end can then be selectively altered to allow ligation between the desired components.

[0124] In addition, as needed, amplification can also optionally be conducted using a wide variety of known techniques to increase the number of genomic fragments for further manipulation, although in many embodiments, an amplification step is not needed at this step.

[0125] After fractionation and optional termini adjustment, a set of adaptor "arms" are added to the termini of the genomic fragments. The two adaptor arms, when ligated together, form the first adaptor. For example, as depicted in Figure 6, circularization (605) of a linear construct with an adaptor arm on each end of the construct ligates the two arms together to form the full adaptor (606) as well as the circular construct (607). Thus, a first adaptor arm (603) of a first adaptor is added to one terminus of the genomic fragment, and a second adaptor arm (604) of a first adaptor is added to the other terminus of the genomic fragment. Generally, and as more fully described below, either or both of the adaptor arms will include a recognition site for a Type IIs endonuclease, depending on the desired system. Alternatively, the adaptor arms can each contain a partial recognition site that is reconstituted upon ligation of the arms.

[0126] In order to ligate subsequent adaptors in a desired position and orientation for sequencing, the present invention provides a method in which a Type IIs restriction endonuclease binds to a recognition site within the first adaptor of a circular nucleic acid construct and then cleaves at a point outside the first adaptor and in the genomic fragment (also referred to herein as the "target nucleic acid"). A second adaptor can then be ligated into the point at which cleavage occurs (again, usually by adding two adaptor arms of the second adaptor). In order to cleave the target nucleic acid at a known point, it can be desirable to block any other recognition sites for that same enzyme that may randomly be encompassed in the target nucleic acid, such that the only point at which that restriction endonuclease can bind is within the first adaptor, thus avoiding undesired cleavage of the constructs. Generally, the recognition site in the first adaptor is first protected from inactivation, and then any other unprotected recognition sites in the construct are inactivated, generally through methylation. That is, methylated recognition sites will not bind the enzyme, and thus no cleavage will occur. Only the unmethylated recognition site within the adaptor will allow binding of the enzyme with subsequent cleaving.

[0127] One method of protecting the recognition site in the first adaptor from inactivation is to make the site single stranded, as the methylation enzyme will not bind to a single strand. Thus, one method of protecting the recognition site of the first adaptor is by amplifying the linear genomic fragments ligated to the two first adaptor arms using primers modified with uracil. The primers are complementary to the adaptor arms and are modified with uracil such that, upon amplification (generally using PCR), the resultant linear constructs contain uracil embedded in the recognition site of one of the first adaptor arms. Digestion of the uracil using known techniques renders that first adaptor arm (or whatever contains the uracil) single stranded. A sequence specific methylase is then applied to the linear constructs that will methylate all of the double-stranded recognition sites for the same endonuclease as that contained in the first adaptor. Such a sequence-specific methylase will not be able to methylate the single stranded recognition site in the first adaptor arm, and thus the recognition site in the first adaptor arm will be protected from inactivation by methylation. As described below, if a restriction site is methylated, it will not be cleaved by the restriction endonuclease enzyme.

[0128] In some cases, as more fully described below, a single adaptor may have two of the same recognition sites, to allow cleavage both "upstream" and "downstream" from the same adaptor. In this embodiment, as depicted in Figure 7, the primers and uracil positions are chosen appropriately, such that either the "upstream" or "downstream" recognition site may be selectively protected from inactivation or inactivated. For example, in Figure 7, the two different adaptor arms (represented as rectangles) each comprise a recognition site for a restriction endonuclease (represented by the circle in one adaptor arm and by a triangle in the other). If the adaptor arm with the recognition site represented by the circle needs to be protected using the above-described uracil degradation method, then the uracil-modified amplification primers are designed to incorporate uracils into that recognition site. Then upon uracil degradation, that adaptor arm is rendered single stranded (represented by the half-rectangles), thus protecting that recognition site from inactivation.

[0129] After protecting the recognition site in the first adaptor arm from methylation, the linear construct is circularized, for example, by using a bridge oligonucleotide and T4 ligase. The circularization reconstitutes the double stranded restriction endonuclease recognition site in the first adaptor arm. In some embodiments, the bridge oligonucleotide has a blocked end, which results in the bridging oligonucleotide serving to allow circularization, ligating the non-blocked end, and leaving a nick near the

recognition site. This nick can be further exploited as discussed below. Application of the restriction endonuclease produces a second linear construct that comprises the first adaptor in the interior of the target nucleic acid and termini comprising (depending on the enzyme) a two base overhang.

[0130] A second set of adaptor arms for a second adaptor is ligated to the second linear construct. In some cases, when a nick is utilized, in order to ensure that the adaptors are ligated in the proper orientation, the nick in the first adaptor is "translated" (or "translocated") by using a polymerase with exonuclease activity. The exonuclease activity of the polymerase (such as Taq polymerase) will excise the short DNA strand that abuts the nick while the polymerase activity will "fill in" the nick and subsequent nucleotides in that strand (essentially, the Taq moves along the strand, excising bases using the exonuclease activity and adding the same bases, with the result being that the nick is translocated along the strand until the enzyme reaches the end).

[0131] In addition, to create an asymmetry of the template, one termini of the construct is modified with a single base. For example, certain polymerases, such as Taq, will undergo untemplated nucleotide addition to result in addition of a single nucleotide to the 3' end of the blunt DNA duplex, resulting in a 3' overhang. As will be appreciated by those in the art, any base can be added, depending on the dNTP concentration in the solution. In certain embodiments, the polymerase utilized will only be able to add a single nucleotide. For example, Taq polymerase will be able to add a single G or A. Other polymerases may also be used to add other nucleotides to produce the overhang. In one embodiment, an excess of dGTP is used, resulting in the untemplated addition of a guanosine at the 3' end of one of the strands. This "G-tail" on the 3' end of the second linear construct results in an asymmetry of the termini, and thus will ligate to a second adaptor arm, which will have a C-tail that will allow the second adaptor arm to anneal to the 3' end of the second linear construct. The adaptor arm meant to ligate to the 5' end will have a C-tail positioned such that it will ligate to the 5' G-tail. After ligation of the second adaptor arms, the construct is circularized to produce a second circular construct comprising two adaptors. The second adaptor will generally contain a recognition site for a Type IIs endonuclease, and this recognition site may be the same or different than the recognition site contained in the first adaptor, with the latter finding use in a variety of applications

[0132] A third adaptor can be inserted on the other side of the first adaptor by cutting with a restriction endonuclease bound to a recognition site in the second arm of the first adaptor (the recognition site that was originally inactivated by methylation). In order to make this recognition site available, uracil-modified primers complementary to the second recognition site in the first adaptor are used to amplify the circular constructs to produce third linear constructs in which the first adaptor comprises uracils embedded in the second restriction recognition site. The uracils are degraded to render the first adaptor single stranded, which protects the recognition site in the adaptor from methylation. Applying a sequence-specific methylase will then inactivate all unprotected recognition sites. Upon circularization the recognition site in the first adaptor is reconstituted, and applying the restriction endonuclease will cleave the circle, producing a position at which the third adaptor can be inserted in a third linear construct. Ligating third adaptor arms to the third linear construct will follow the same general procedure described above – the third linear construct will be A- or G-tailed, the third adaptor arms will be T- or C-tailed, allowing the adaptor arms to anneal to the third linear construct and be ligated. The linear construct comprising the third adaptor arms is then circularized to form a third circular construct. Like the second adaptor, the

third adaptor will generally comprise a recognition site for a restriction endonuclease that is different than the recognition site contained in the first adaptor.

[0133] A fourth adaptor can be added by utilizing Type IIs restriction endonucleases that have recognition sites in the second and third adaptors. Cleavage with these restriction endonucleases will result in a fourth linear construct that can then be ligated to fourth adaptor arms. Circularization of the fourth linear construct ligated to the fourth adaptor arms will produce the nucleic acid template constructs of the invention. As will be appreciated by those in the art, other adaptors can be added. Thus, the methods described herein allow two or more adaptors to be added in an orientation and sometimes distance dependent manner.

[0134] The present invention also provides methods for controlling the orientation in which each subsequently added adaptor is inserted. Such "nick translation" methods provide a way to control the way target nucleic acids and adaptors ligate to each other. These methods also prevent artifacts in the nucleic acid constructs by preventing ligation of adaptors to other adaptors and target nucleic acid molecules to other target nucleic acid molecules (essentially avoiding the "polymerization" of adaptors and target nucleic acid molecules). Examples of different orientations in which adaptors and target nucleic acid molecules can be ligated are schematically illustrated in Figure 8. Target nucleic acids 801 and 802 are preferably ligated to adaptors 803 and 804 in a desired orientation (as illustrated in this figure, the desired orientation is one in which the ends with the same shape –circle or square – ligates to each other). Modifying the ends of the molecules avoids the undesired configurations 807, 808, 809 and 810, in which the target nucleic acids ligate to each other and the adaptors ligate to each other. In addition, as will be discussed in further detail below, the orientation of each adaptor-target nucleic acid ligation can also be controlled through control of the chemistry of the termini of both the adaptors and the target nucleic acids. The control over the chemistry of the termini can be provided using methods known in the art. For example, in some circumstances, the use of phosphatase eliminates all the phosphate groups, such that all ends contain hydroxyl groups. Each end can then be selectively altered to allow ligation between the desired components. These and other methods for modifying ends and controlling insertion of adaptors in the nick translation methods of the invention are described in further detail below.

[0135] These nucleic acid template constructs ("monomers" comprising target sequences interspersed with these adaptors) can then be used in the generation of concatemers, which in turn form the nucleic acid nanoballs that can be used in downstream applications, such as sequencing and detection of specific target sequences.

[0136] The present invention provides methods for forming nucleic acid template constructs comprising multiple interspersed adaptors inserted into a target nucleic acid. As discussed further herein, methods of the invention allow insertion of each subsequent adaptor by utilizing recognition sites for Type IIs restriction endonucleases that are included in the adaptors. In order to insert multiple adaptors in a desired order and/or orientation, it can be necessary to block restriction endonuclease recognition sites contained within the target nucleic acids, such that only the recognition site in the adaptor is available for binding the enzyme and the subsequent cleavage. Among the advantages of such methods is that the same restriction endonuclease site can be used in each adaptor, which simplifies production of circular templates that will eventually be used to generate concatemers, adaptors can be inserted using a previously inserted adaptor as a "stepping stone" for the next, such that addition can occur in effect by "walking" down the length of the fragment with each new adaptor. Controlling the recognition sites

available for restriction enzymes also avoids the excision of certain sequences, thereby obtaining only limited sequence representation (which could result if sites within the target nucleic acid were accessible).

IVB. Adding a first adaptor

[0137] As a first step in the creation of nucleic acid templates of the invention, a first adaptor is ligated to a target nucleic acid. The entire first adaptor may be added to one terminus, or two portions of the first adaptor, referred to herein as "adaptor arms", can be ligated to each terminus of the target nucleic acid. The first adaptor arms are designed such that upon ligation they reconstitute the entire first adaptor. As described further above, the first adaptor will generally comprise one or more recognition sites for a Type IIs restriction endonuclease. In some embodiments, a Type IIs restriction endonuclease recognition site will be split between the two adaptor arms, such that the site is only available for binding to a restriction endonuclease upon ligation of the two adaptor arms.

[0138] Figure 6 is a schematic representation of one aspect of a method for assembling adaptor/target nucleic acid templates (also referred to herein as "target library constructs", "library constructs" and all grammatical equivalents). DNA, such as genomic DNA 601, is isolated and fragmented into target nucleic acids 602 using standard techniques as described above. The fragmented target nucleic acids 602 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. The A-tailing is generally accomplished by using a polymerase (such as Taq polymerase) and providing only adenosine nucleotides, such that the polymerase is forced to add one or more A's to the end of the target nucleic acid in a template-sequence-independent manner.

[0139] In the exemplary method illustrated in Figure 6, a first (603) and second arm (603) of a first adaptor is then ligated to each target nucleic acid, producing a target nucleic acid with adaptor arms ligated to each end. In one embodiment, the adaptor arms are "T tailed" to be complementary to the A tails of the target nucleic acid, facilitating ligation of the adaptor arms to the target nucleic acid by providing a way for the adaptor arms to first anneal to the target nucleic acids and then applying a ligase to join the adaptor arms to the target nucleic acid.

[0140] In a further embodiment, the invention provides adaptor ligation to each fragment in a manner that minimizes the creation of intra- or intermolecular ligation artifacts. This is desirable because random fragments of target nucleic acids forming ligation artifacts with one another create false proximal genomic relationships between target nucleic acid fragments, complicating the sequence alignment process. 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 artifacts that would be created from self-ligation, adaptor-adaptor or fragment-fragment ligation.

[0141] As an alternative to A/T tailing (or G/C tailing), various other methods can be implemented to prevent formation of ligation artifacts 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.

[0142] After creating a linear construct comprising a target nucleic acid and with an adaptor arm on each terminus, the linear target nucleic acid is circularized (605), a process that will be discussed in further detail herein, resulting in a circular construct 607 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 first adaptor (606) in the circular construct. In some embodiments, the circular construct 607 is amplified, such as by circle dependent amplification, using, *e.g.,* random hexamers and Φ29 or helicase. Alternatively, target nucleic acid/adaptor structure may remain linear, and amplification may be accomplished by PCR primed from sites in the adaptor arms. The amplification 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.

IVC. Adding multiple adaptors

[0143] Figure 6 is a schematic representation of one aspect of a method for assembling adaptor/target nucleic acid templates (also referred to herein as "target library constructs", "library constructs" and all grammatical equivalents). DNA, such as genomic DNA 601, is isolated and fragmented into target nucleic acids 102 using standard techniques. The fragmented target nucleic acids 602 are then in some embodiments (as described herein) repaired so that the 5' and 3' ends of each strand are flush or blunt ended.

[0144] In the exemplary method illustrated in Figure 6, a first (603) and second arm (604) of a first adaptor is ligated to each target nucleic acid, producing a target nucleic acid with adaptor arms ligated to each end.

[0145] After creating a linear construct comprising a target nucleic acid and with an adaptor arm on each terminus, the linear target nucleic acid is circularized (605), a process that will be discussed in further detail herein, resulting in a circular construct 607 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 first adaptor (606) in the circular construct. In some embodiments, the circular construct 607 is amplified, such as by circle dependent amplification, using, *e.g.,* random hexamers and Φ29 or helicase. Alternatively, target nucleic acid/adaptor structure may remain linear, and amplification may be accomplished by PCR primed from sites in the adaptor arms. The amplification 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.

[0146] Similar to the process for adding the first adaptor, a second set of adaptor arms (610) and (611) can be added to each end of the linear molecule (609) and then ligated (612) to form the full adaptor (614) and circular molecule (613). Again, a third adaptor can be added to the other side of adaptor (609) by utilizing a Type IIs endonuclease that cleaves on the other side of adaptor (609) and then ligating a third set of adaptor arms (617) and (618) to each terminus of the linearized molecule. Finally, a fourth adaptor can be added by again cleaving the circular construct and adding a fourth set of adaptor arms to the linearized construct. The embodiment pictured in Figure 6 is a method in which Type IIs endonucleases with recognition sites in adaptors (620) and (614) are applied to cleave the circular

construct. The recognition sites in adaptors (620) and (614) may be identical or different. Similarly, the recognition sites in all of the adaptors illustrated in Figure 6 may be identical or different.

[0147] As generally illustrated in Figure 9, a circular construct comprising a first adaptor may contain two Type IIs restriction endonuclease recognition sites in that adaptor, positioned such that the target nucleic acid outside the recognition sequence (and outside of the adaptor) is cut (910). The arrows around structure 510 indicate the recognition sites and the site of restriction. In process 911, EcoP15, a Type IIs restriction endonuclease, is used to cut the circular construct. Note that in the aspect shown in Figure 9, 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 910). Restriction of the library constructs with EcoP15 in process 911 results in a library of linear constructs containing the first adaptor, with the first adaptor "interior" to the ends of the linear construct 912. 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 913, the linear construct 912, like the fragmented target nucleic acid 904, 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 913. The resulting library construct comprises the structure seen at 914, 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.

[0148] In process 915, the double-stranded linear library constructs are treated so as to become single-stranded 916, and the single-stranded library constructs 916 are then ligated 917 to form single-stranded circles of target nucleic acid interspersed with two adaptors 918. The ligation/circularization process of 917 is performed under conditions that optimize intramolecular ligation. At certain concentrations and reaction conditions, the local intramolecular ligation of the ends of each nucleic acid construct is favored over ligation between molecules.

IVD. Controlling orientation of ligation between target nucleic acids and adaptors

[0149] In one aspect, the present invention provides methods in which ligation of adaptors to target nucleic acids, as described above, is accomplished in a desired orientation. Such control over orientation is advantageous, because random fragments of target nucleic acids forming ligation artifacts with one another create false proximal genomic relationships between target nucleic acid fragments, complicating the sequence alignment process.

[0150] There are several methods that find use in controlling orientation of the adaptor insertion. As described above, altering the chemistry of the termini of the targets and the adaptors can be done, such that ligation can only occur when the correct orientation is present. Alternatively, "nick translation methods" can be done, which also rely on the termini chemistries, as outlined below. Finally, methods involving amplification with specific choices of primers can be done as described below.

[0151] Figure 12 is a schematic illustration of the different orientations in which a second adaptor may be added to a nucleic acid construct. Again, process 1200 begins with circular library construct 1202, having an inserted first adaptor 1210. First adaptor 1210 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 (Ad1 orientation 1210). A Type IIs restriction endonuclease site in the first adaptor 1210 is indicated by the tail of arrow 1201, and the site of cutting is indicated by the arrow head. Process 1203 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 1204 and 1206, the second adaptor can be inserted in two different ways relative to the first adaptor. In the desired orientation 1204, the oval is inserted into the outer strand with the rectangle, and the bowtie is inserted into the inner strand with the diamond (Ad2 orientation 1220). In the undesired orientation the oval is inserted into the inner strand with the diamond and the bowtie is inserted into the outer strand with the rectangle (Ad2 orientation 1230).

[0152] Although much of the following discussion and referenced illustrative figures discuss for clarity's sake insertion of a second adaptor in relation to a first, it will be appreciated that the processes discussed herein are applicable to adaptors added subsequently to the second adaptor, creating library constructs with three, four, five, six, seven, eight, nine, ten or more inserted adaptors.

[0153] In one embodiment, both A tailing and T tailing are used to attach an adaptor to a nucleic acid fragment. For example, following the modifications described above to repair the ends of fragments, each fragment can be "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. The A-tailing is generally accomplished by using a polymerase (such as Taq polymerase) and providing either only adenosine nucleotides (or an excess thereof), such that the polymerase is forced to add one or more A's to the end of the target nucleic acid in a template-sequence-independent manner. In embodiments in which "A-tailing" is used, ligation to adaptor (or adaptor arms) can be accomplished by adding a "T-tail" to the 5' end of the adaptor/adaptor arms to be complementary to the A tails of the target nucleic acid, facilitating ligation of the adaptor arms to the target nucleic acid by providing a way for the adaptor arms to first anneal to the target nucleic acids and then applying a ligase to join the adaptor arms to the target nucleic acid.

[0154] Because the aspects of the claimed invention work optimally when nucleic acid templates are of a desired size and comprise target nucleic acid derived from a single fragment, it can be beneficial to ensure that throughout the process of producing nucleic acid templates that the circularization reactions occur intramolecularly. That is, it can be beneficial to ensure that target nucleic acids in the process of being ligated to a first, second, third, etc. adaptor do not ligate to one another. One embodiment of controlling the circularization process is illustrated in Figure 10. As shown in Figure 10, blocking oligos 1017 and 1027 are used to block the binding regions 1012 and 1022 regions, respectively. Blocker oligonucleotide 1017 is complementary to binding sequence 1016, and blocker oligonucleotide 1027 is complementary to binding sequence 1026. In the schematic illustrations of the 5' adaptor arm and the 3' adaptor arm, the underlined bases are dideoxycytosine (ddC) and the bolded font bases are phosphorylated. Blocker oligonucleotides 1017 and 1027 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 ligation of the adaptor arms (circularization).

IVD(i). Controlling orientation of ligation: arm-by-arm ligation

[0155] In one aspect, the directional insertion of adaptors can be controlled without modifying the termini of the target nucleic acid using an "arm-by-arm" ligation method. In general, this is a two-step ligation process in which an adaptor arm is added to a target nucleic acid and primer extension with strand displacement produces two double stranded molecules each with an adaptor arm on one end – a second adaptor arm can then be ligated to the terminus without an adaptor arm. This process can prevent the creation of nucleic acid molecules that comprise the same adaptor arm on both termini – for example, as depicted in Figure 11A, the arm-by-arm ligation process can prevent the formation of nucleic acid molecules that have both termini occupied by Adaptor A or Adaptor B. In many embodiments it is preferred that each terminus of a target nucleic acid is ligated to a different adaptor arm, such that when the two arms are ligated they are able to form a complete whole adaptor. This can be particularly useful for minimizing the number of amplification steps that are needed after addition of each adaptor arm, because the arm-by-arm ligation reduces the number of non-useful molecules produced in each ligation reaction.

[0156] Figure 11 illustrates one embodiment of the arm-by-arm ligation method. In this embodiment, one strand of the first adaptor arm A is added to both strands of a dephosphorylated target nucleic acid. This adaptor arm is blocked on one end (depicted as the closed circle), generally by using alkaline phosphatase. Primer exchange can be used to replace the strand with the blocked end. Primer extension with strand displacement (which can be accomplished, in one exemplary embodiment, through the use of phi29 or Pfu polymerase) will prime from both ends and extend through the whole insert, resulting in two double-stranded nucleic acid molecules, each with an adaptor arm A on one terminus and a blunt end on the other. In an alternative embodiment, adaptor arm A can be used pre-hybridized with a primer upstream of the blocked strand to initiate primer extension without requiring a primer exchange reaction. After the strand-displacing polymerase reaction, a second adaptor arm B can then be ligated, generally to the blunt end of the target nucleic acid rather than to the terminus with the adaptor arm. This arm-by-arm ligation process can prevent the formation of target nucleic acids that comprise the same adaptor arm on both termini.

IVD(ii). Controlling orientation of ligation: nick translation methods

[0157] In one embodiment, the present invention provides "nick translation methods" for constructing nucleic acid molecules. In one embodiment, nick translation methods are used to ligate nucleic acid molecules in a desired orientation. In a further embodiment, nick translation methods are used for inserting adaptors in a desired orientation. Such methods generally involve modifying one or both termini of one or both of the nucleic acid molecules to be ligated together. For example, when ligating an adaptor to a target nucleic acid, one or both termini of either or both the target nucleic acid and adaptor to be ligated are modified. Following such modification, a "translocation" or "translation" of a nick inserted into one strand of a construct provides the ability to control the final orientation of the ligated adaptor-target nucleic acid construct. "Nick translation methods" as described herein may also include

primer extension or gap-fill-in methods, as is described in further detail below. Although the following discussion is provided in terms of controlling ligation of adaptors to target nucleic acids, it will be appreciated that these methods are not limited to ligation of adaptors and target nucleic acids, and that these methods can also be used to control ligation of any two nucleic acid molecules. For example, nick translation methods and any other controlled ligation methods described herein can be used as part of genetic and/or DNA engineering methods, such as the construction of new plasmids or other DNA vectors, gene or genome synthesis or modifications, as well as in constructing building blocks for nanotechnology constructs.

[0158] Figure 13 is a schematic illustration of such a "nick translation" type of process. Construct 1306 in Figure 13 is formed using methods discussed herein, and has an interspersed adaptor 1304, with a restriction endonuclease recognition site (tail of the arrow in Figure 13), and a cleavage site. In Figure 14, the library construct is not circularized, but is a branched concatemer of alternating target nucleic acid fragments 1406 (with restriction endonuclease recognition sites 1404) and adaptors 1412; however, the nick translation type process shown in Figure 13 may be performed on such a library construct configuration as well. The term "library construct" as used herein refers to nucleic acid constructs comprising one or more adaptors, and is interchangeable with the term "nucleic acid template".

[0159] The library constructs with an inserted first adaptor are digested by a restriction endonuclease (process 1301)--in certain aspects, a Type IIs restriction endonuclease--that cuts the target nucleic acid to render 3' nucleotide overhangs 1308. In Figure 11, two nucleotides (NN-3') 1308 are shown, though the number of overhanging nucleotides varies in alternative aspects depending at least in part on the identify of the restriction endonuclease used. The library construct 1310 is linearized, with the first inserted adaptor shown at 1304. The first inserted adaptor 1304 is engineered such that it comprises either a nick 1312 at the boundary of the adaptor fragment or it comprises the recognition site for a nicking endonuclease that permits the introduction of a nick 1314 at the interior of the adaptor. In either case, library construct 1310 is treated 1303 with a polymerase 1316 that can extend the upper strand from nick 1312 or 1314 to the end of the lower strand of library construct 1310 to form a strand having a 3' overhang at one end and a blunt end at the other. To this library construct 1310, a second adaptor 1318 is ligated in process 1305, where the second adaptor 1318 has a degenerate nucleotide overhang at one end and a single 3' nucleotide (*e.g.*, dT) overhang at the other end to form library construct 1320. Library construct 1320 is then treated (*e.g.*, with Taq polymerase) in process 1307 to add a 3' dA to the blunt end. Library construct 1322 may then be amplified by PCR, with, *e.g.*, uracil-containing primers. Alternatively, library construct 1322 may then be circularized in process 1309 in which case CDA may be performed (such as in step 1421 of Figure 14). Combining the processes discussed herein with the nick translation type process shown in Figure 13 allows for selecting both the relative position and relative orientation of subsequently-added adaptors to any adaptors previously inserted into the library constructs.

[0160] In order to utilize a nick translation type of procedure, it may be beneficial to modify one or both of the termini of the target nucleic acid and/or the adaptor as discussed above. In one exemplary embodiment, a first arm of an adaptor that is meant to ligate to the 3' end of a target nucleic acid may be designed such that its 3' terminus is blocked, such that only the 5' end of the adaptor arm is available to ligate to only the 3' end of the target nucleic acid. Similarly, the second arm that is meant to ligate to the 5' end of the target nucleic acid may be designed such that its 5' end is blocked, such that only the 3' end

of the second arm can ligate to the 5' end of the target nucleic acid. Methods for blocking one terminus of the adaptor arm and/or the target nucleic acid are well known in the art. For example, the target nucleic acid (which is also referred to herein as a "nucleic acid insert" or a "DNA insert" or an "insert") is treated with enzymes that generate defined functional ends and remove phosphates from both the 3' and 5' ends as discussed above. Removing all of the phosphate groups renders the target nucleic acid molecules unable to ligate to each other. Adaptors in this embodiment are also designed to have one strand capable of ligation (for example by creating or maintaining a 5' phosphate group) and a complementary strand that has a 3' end that is protected from ligation. Generally, this protection of the 3' end is accomplished using a dideoxy nucleotide to inactivate the 3' end. Thus, when the modified target nucleic acids lacking phosphate groups on both ends and modified adaptors comprising only a phosphate group on one 5' end with a 3' block (for example, a dideoxy) on the complementary strand, the only ligation product that will form is that of target nucleic acid ligated to the 5' end of the adaptor that has a phosphate group. Subsequent to this ligation step, the protected 3' end of the adaptor can be exchanged with a strand containing a functional 3' end. This exchange is generally accomplished by taking advantage of the fact that the strand with 3' protection is generally short and easy to denature. The exchange strand with a functional 3' end is longer and will thus bind more efficiently to the complementary strand – in further embodiments, the strand with the functional end is also added in higher concentrations to further influence the reaction toward exchanging the protected strand with the strand with the functional end. This strand with the functional 3' end is then primed by adding a DNA polymerase with nick translation activity, such that the polymerase exonucleolytically removes bases from the 5' end of the target nucleic acid, thereby exposing a functional 5' phosphate. This newly generated 5' phosphate can be ligated to the extension product by a ligase. (If ligase is absent during the extension reaction, two polymerase molecules will nick translate from each end of the target nucleic acid until they meet each other, resulting in a broken molecule). For example, as illustrated in Figure 2, the target nucleic acid (insert) is first end-repaired to form defined functional ends, preferentially blunt-ends. Next, to avoid concatemerization of inserts, 5'-end phosphates are removed. The insert is then mixed with DNA ligase and DNA adaptors. The DNA adaptor contains two oligonucleotides, and has one blunt-end and one sticky-end when the two oligonucleotides are hybridized together. The blunt-end side contains one "top-strand" with a protected/inactivated 3'-end, and one "bottom-strand" with a functional 5'-end phosphate, and are thus also unable to self-ligate. The only possible ligation combination is therefore one insert with one "bottom-strand" blunt-ligated to each end. The "top-strand" with 3'-end protection is then exchanged with an oligonucleotide containing a functional 3'-end that can act as a primer in a polymerase extension reaction. Upon addition of polymerase and ligase, the second oligonucleotide can be built-in through a nick translation and ligation reaction. When the polymerase is extending into the insert, it introduces a nick with a functional 5'-end phosphate that can be recognized and sealed by DNA ligase. The resulting insert with an adaptor or adaptor arm on each side of each strand can now be subjected to PCR using primers specific to the adaptor.

[0161] Generally in a nick translation reaction such as the one described above, an active ligase is present or added in the mixture before addition of the polymerase or simultaneously with the polymerase. In some embodiments, it can be beneficial to use low activity polymerase (slow nick translation) conditions. Both addition of the ligase before or simultaneously with the polymerase and low activity conditions can help assure that the translating nick is sealed before reached the opposite end of the DNA

fragment. In some embodiments, this can achieved by incubating the Taq polymerase with T4 ligase at 37°C, a temperature that will usually result in low polymerase activity and high ligase activity. The reaction may then be further incubated at a higher temperature (such as 50-60°C) to further assure nick-translation-ligation occurs to completion across most/all constructs in the reaction.

[0162] In further embodiments, the present invention provides methods for forming nucleic acid template constructs comprising multiple interspersed adaptors. Methods of the present invention include methods of inserting multiple adaptors such that each subsequent adaptor is inserted in a defined position with respect to one or more previously added adaptors. Certain methods of inserting multiple interspersed adaptors are known in the art, for example, as discussed in U.S. Application Serial Nos. 60/992,485; 61/026,337; 61/035,914; 61/061,134; 61/116,193; 61/102,586; 12/265,593; 12/266,385; 11/679,124; 11/981,761; 11/981,661; 11/981,605; 11/981,793 and 11/981,804, each of which is herein incorporated by reference in its entirety for all purposes and in particular for all teachings related to methods and compositions for creating nucleic acid templates comprising multiple interspersed adaptors as well as all methods for using such nucleic acid templates. Insertion of known adaptor sequences into target sequences, such that there is an interruption of contiguous target sequence with the multiple interspersed adaptors, provides the ability to sequence both “upstream” and “downstream” of each adaptor, thus increasing the amount of sequence information that can be generated from each nucleic acid template. The present invention provides further methods for inserting each subsequent adaptor in a defined position with respect to one or more previously added adaptors.

[0163] Nick translation ligation is usually performed after ligating the first strand by adding at least polymerase to the reaction. In some embodiments, the nick translation reaction may be performed as a one-step reaction by adding all components at once, while in some embodiments the steps of the reaction are performed sequentially. There are multiple possible embodiments of a “one-step” approach of the nick translation reaction. For example, a single mix with a primer can be used in which Taq is added at the beginning of the reaction. Use of a thermo-stable ligase provides the ability of performing primer exchange and nick translation ligation (and PCR if necessary) by simply increasing the temperature. In another exemplary embodiment, the reaction mixture will contain a minimal concentration of non-processive nick-translating polymerase with a weak 3’ exonuclease that activates the 3’ blocked strand.

[0164] In a further embodiment, T4 polynucleotide kinase (PNK) or alkaline phosphatase is used to alter 3’ ends of adaptors and/or target nucleic acids to prepare them for a nick translation process. For example, adaptors can be inserted as part of a circularization reaction. End-repaired and alkaline phosphatase treated target nucleic acids are ligated to adaptors, and in this exemplary embodiment are designed to form self-complementary hairpin shaped units (Figure 16). The hairpins are designed to contain modifications at a given position that can be recognized and cleaved by enzymes or chemicals. For example, if the hairpins contain deoxyuridines, the deoxyuridiines can be recognized and cut by UDG/EndoVIII. After cutting, the two hairpins become single-stranded with phosphates on their respective 3-end. These 3’ phosphates can then be removed by either T4 Polynucleotide Kinase (PNK) or alkaline phosphatase (SAP) to enable nick-translation methods as described further herein. In an exemplary embodiment such as the one illustrated in Figure 4A, the two hairpins are designed to be partly complementary to each other and can thus form, by intra-molecular hybridization, circularized molecules. Finally, the circularized molecules are subjected to a nick-translation process in which a

polymerase extends into the insert and introduces a nick with a functional 5'-end phosphate that can be recognized and sealed by DNA ligase.

[0165] Instead of using hairpins as described above, a pair of double stranded adaptors that are partly complementary to each other can be used for circularization. One pair has deoxyuridines on one strand that can be recognized and cut by UDG/EndoVIII. Other methods of nicking one strand can also be used, including without limitation: nicking enzymes, incorporating inosine modified DNA that can be recognized by endonucleolytic enzymes, and incorporating DNA with RNA modifications that can be recognized by RNA-endonucleases. The target nucleic acid and adaptors can be prepared for controlled ligation as described above, for example by treating the target nucleic acid with alkaline phosphatase to create blunt ends that are unable to ligate to other target nucleic acid. Circularization is activated by denaturing the short 3'-protected strand in the adaptor from the strand ligated to the target nucleic acid, leaving two partly complementary single stranded ends on each end of the target nucleic acid insert. The ends are then joined by intra-molecular hybridization and subjected to nick-translation and ligation, forming a covalently closed circle. The circles are then treated with UDG/EndoVIII to prepare the circle for directional insertion of the next adaptor.

[0166] In a still further embodiment illustrated in Figure 15, a linear target nucleic acid is treated with shrimp alkaline phosphatase (SAP) to remove 5' phosphates. Next, the target nucleic acid is ligated to one arm of the adaptor (arm A), containing a strand with a 5' phosphate, and a complementary shorter strand with a protected 3' end. The ligation product is then subjected to nick-translation. The nick generated in the circularization reaction is located on the top strand of the first adaptor, and acts as a primer for the polymerase used in the nick-translation reaction. The polymerase extends the top-strand to the nick at the adaptor-insert junction, releasing one of the adaptor A arms and generating blunt end or A or G overhang . Next, the resulting polymerase-generated insert end is ligated to the second adaptor arm (arm B). By designing the first adaptor to generate a nick in the circularization reaction, the subsequent adaptor can be added in a predetermined orientation. This strategy is applicable for all type IIs restriction enzymes or other enzymatic or non-enzymatic fragmenting methods regardless of whether they generate a digested product that has blunt ends, 3' overhangs, or 5' overhangs. The subsequent primer exchange, extension, ligation, and PCR is similar to that described in figure 2. A no-amplification option may also be used to close the circle comprising melting off the blocked oligonucleotides followed by DNA circularization via nick translation ligation reaction.

[0167] Both proofreading polymerases (which have 3'-5' exonuclease activity), such as Pfu polymerase, and non-proofreading polymerases (which lack 3'-5' exonuclease activity), such as *Taq* polymerase, can be used in the nick translation and strand synthesis with strand displacement processes described herein. Proofreading polymerases can efficiently generate blunt-ends in the nick translation process but have the disadvantage of also degrading non-protected 3' overhangs. The resulting nick translation product will therefore have two blunt ends and will thus be unable to ligate subsequent adaptors in defined orientation. One solution is to protect the 3' end of the ligated adaptor (arm A in Figure 15 for example) from degradation, using e.g. dideoxyribonucleoside triphosphates (ddNTP) on the 3' ends. However, ddNTP protection also protects the 3' end from subsequent extension, thus limiting the adaptors to be carried forward in a direct circularization procedure. Another potential solution is to protect the 3' ends from polymerase degradation using modifications on the 3' end (e.g. 3' phosphate) that can be removed prior to nick translation circularization (e.g. using alkaline phosphatase). Another

approach is to use hairpin shaped adaptors (as described in figure 16) in combination with proofreading polymerase in nick translation reactions. These adaptors will be protected from degradation but have the disadvantage of requiring an extra UDG/EndoVIII step. Furthermore, the inventors have found that one of the proofreading polymerases, Pfu polymerase, is able to efficiently generate blunt ends without degrading the non-protected 3' overhang, indicating a low 3'-5' exonuclease activity.

[0168] Non-proofreading polymerases, such as Taq polymerase, can generate both blunt ends and single base overhangs in the nick translation process (Taq can generate non-templated A- and G-tails in addition to blunt ends). An advantage of using polymerases without 3'-5' exonuclease activity in the nick translation process is that non-protected 3' overhangs remain intact. This enables ligation of subsequent adaptors in defined orientation without protecting 3' overhangs from degradation. A potential disadvantage with many proofreading polymerases is that they have a function of adding single nucleotides on 3' ends in a non-templated process. This process can be hard to control, and will often generate a mixed population of 3' ends, resulting in a low adaptor-to-insert ligation yield. In general, methods utilizing blunt end ligation are more efficient than one base overhang ligation.

[0169] In one embodiment, after ligation of a first adaptor, rather than forming a circle and then cleaving with a type IIs endonuclease that has a recognition site in the first adaptor (which is a step in some embodiments of producing nucleic acid templates of the invention, such as embodiments schematically illustrated in Figures 6 and 9), a second adaptor can be added using a variation of the nick translation method. Exemplary embodiments of this variation are schematically illustrated in Figure 17. In general, these embodiments begin with addition of a first adaptor to a target nucleic acid and then circularization, as is described in detail above and illustrated in Figures 6 and 9. In the embodiment illustrated in Figure 17A, a nick translation is carried out using a polymerase with 5'-3' exonuclease activity (such as Taq polymerase), which generates an inverted circle with the first adaptor located in the interior of the target nucleic acid. This product can then be end-repaired and subjected to ligation to adaptor 2 (using methods described in further detail above). One disadvantage of this embodiment is that the target nucleic acid may be longer than is required for sequencing application, and such longer templates might be prone to generating secondary structures in any nucleic acid concatemer products generated from the templates (the generation of concatemers from nucleic acid templates of the invention is discussed in greater detail below). Such secondary structures may result in a decreased signal when these concatemers are used in sequencing applications, such as the cPAL methods discussed below. One way to overcome this disadvantage is by shortening the target nucleic acid – one exemplary embodiment of this approach is pictured in Figure 17B. In this embodiment, the first adaptor is modified with uracils using methods described herein. Following the nick translation-inversion of the circle comprising the first adaptor, an adaptor C-arm is added to both ends of the end-repaired molecule. The uracil-modified adaptor 1 is treated to remove the uracils, creating gaps, and also treated to generate activated 3' ends. Generally, the uracils are removed by application of an UDG/EndoVIII enzyme mix and PNK and/or alkaline phosphatase is used to remove the 3' phosphates and generate activated 3' ends. The activated 3' ends of the adaptor 1 and the 3' ends of adaptor arm C are recognized by a nick translation polymerase (i.e., a polymerase with 5'-3' exonuclease activity) resulting in a product with adaptor 1 surrounded by a target nucleic acid that has been trimmed to approximately half of its original length. This polymerase cutting procedure can be repeated to decrease the size of the target nucleic acid even

further if adaptor 1 is modified with additional nicking modifications (including without limitation incorporation of inosine, RNA-modifications, and the like).

[0170] In a further embodiment, as is illustrated in Figure 17C, the nick translation methods illustrated in Figures 17A and B can be expanded to insert multiple adaptors. By modifying adaptors, nicks or gaps and functional 3' ends can be generated to prime nick translation reactions from multiple adaptors simultaneously. As illustrated in Figure 17C, a nucleic acid construct comprising target nucleic acid and two adaptors, each containing a uracil modification on one strand, is circularized. Next, the circle is treated with an enzyme mix, such as UDG/EndoVIII, to remove the uracils and introduce gaps. These gaps can be simultaneously nick translated to invert the circle, making the construct available for ligation to additional adaptors. By adding multiple modifications on the same adaptors, subsequent nicking/gapping and nick translation inversion can be carried out to introduce multiple adaptors. In some embodiments, uracils can be added back to the same positions in the adaptors, making the adaptors suitable for further nick translation reactions. Adding the uracils back can be accomplished, for example, by incubating the nick translation reaction with uracil only to "build back" the modification in the adaptor, followed by addition of non-modified nucleotides in higher concentration to fill in the rest of the construct.

[0171] In a still further embodiment, illustrated in Figure 17D, the target nucleic acid may be trimmed by controlling the speed of the nick translation enzyme. For example, the nick translation enzyme can be slowed by altering the temperature or limiting reagents, which can result in two nicks being introduced into the circularized insert that are shifted from the initial sites in the adaptor using a nick translation process. Similarly, using a strand displacement polymerase (such as phi29) will result in a nick being shifted, producing a branching point due to a displaced segment of the nucleic acid. These nick or branch points can be recognized by various enzymes (including without limitation S1 endonuclease, Bal31, T7 endonculease, Mung Bean endonuclease, as well as combinations of enzymes, such as a 5' to 3' exonuclease such as T7 exonuclease and S1 or Mung Bean endonuclease) that will cut the opposite strand of the nick, resulting in a linear product. This product can then be end-repaired (if needed) and then ligated to the next adaptor. The size of the target nucleic acid remaining will be controlled by the speed of the nick translation reaction, again for example by lowering the concentration of reagents such as dNTPs or by conducting the reaction at a less than optimal temperature. The size of the target nucleic acid may also be controlled by the incubation time of the nick translation reaction.

[0172] In a further embodiment, nick translation methods can be used to form nucleic acid templates without transitioning through any circularizing steps. An exemplary embodiment of such methods is illustrated in Figure 18, which shows that the first adaptor 1801, which is shaped as a hairpin, is ligated to target nucleic acid 1802 using ligation methods described above, such as by treating the target nucleic acid with shrimp alkaline phosphate to remove phosphate groups and thereby control the ends of the target nucleic acid that are available to ligate to the first adaptor. After ligation of the first adaptor, a controlled double-strand specific 5'-3' exonuclease reaction is carried out to generate single stranded 3' ends. In some embodiments, the exonuclease reaction is carried out using a T7 exonuclease, although it will be appreciated that other double-strand specific exonucleases can be used in this embodiment of the invention. In further embodiments, the exonuclease reaction generates single stranded 3' ends of about 100 to about 3000 bases in length. In still further embodiments, the exonuclease reaction generates single stranded 3' ends of about 150 to about 2500, about 200 to about 2000, about 250 to about 1500,

about 300 to about 1000, about 350 to about 900, about 400 to about 800, about 450 to about 700, and about 500 to about 600 bases in length.

[0173] It will be appreciated that the nick translation processes described herein can be used in combination with any of the other methods of adding adaptors described herein. For example, the arm-by-arm ligation process described above and schematically illustrated in Figure 11A can be used in combination with a nick translation process to prepare a construct for PCR amplification.

[0174] In a further embodiment, adaptor arm A used in an arm-by-arm ligation reaction can be designed for direct circularization without PCR, followed by nick translation ligation to seal the circle. In an exemplary embodiment, for direct circularization, adaptor arm A can be designed as pictured in Figure 11B. Segment 1101 is designed to be complementary to adaptor arm B. The construct in Figure 11B allows for direct primer extension by a strand displacing polymerase (such as phi29) without a need for a primer exchange reaction to remove a blocked end (the polymerase will not extend past the 3' phosphate on segment 1102). This construct also provides a 3' overhang for circularization. Segment 1102 prevents hybridization of adaptor arm A to adaptor arm B before circularization. In some embodiments, segment 1102 may not be necessary for preventing hybridization to arm B (such as when adaptor arm B is provided in very high concentrations) or segment 1102 may be part of the design of adaptor arm B rather than adaptor arm A.

[0175] After generating the single stranded 3' ends, a second adaptor 1803 is hybridized to the single stranded 3' end of the target nucleic acid and connected to the first adaptor through a nick translation ligation reaction (in one embodiment, the nick translation ligation is a "primer extension" or "gap fill-in" reaction). The second adaptor has a 5' phosphate and a 3' block (identified as the vertical line 1804). The 3' block can in some embodiments be a removable block such as a 3' phosphate, which can be removed in some exemplary embodiments using polynucleotide kinase (PNK) and/or shrimp alkaline phosphate. The second adaptor may in some embodiments have degenerated bases at the 3' and/or the 5' ends. In some exemplary embodiments, the second adaptor has about 2-6 degenerated bases at the 5' end and 4-9 degenerated bases at the 3' end, although it will be appreciated that any combination of numbers of degenerated bases at one or both ends of the second adaptor are encompassed by the present invention. In the embodiment pictured in Figure 18, the second adaptor comprises 3 degenerate bases at the 5' end ("N3") and 7 degenerate bases at its 3' end ("N7"). The joining of the first adaptor to the second adaptor may in some embodiments be accomplished under reaction conditions at which hybridization of the adaptors to the target nucleic acid are favored. In some exemplary embodiments, such reaction conditions may include temperatures of from about 20 to about 40°C. Polymerases that can be used under such reaction conditions include without limitations phi29, Klenow, T4 polymerases and Pol I.

[0176] The ligation product 1805 is then denatured and/or further processed with a 5'-3' exonucleases followed by a re-annealing step to form two single stranded nucleic acid molecules (denoted by the "x2" in Figure 18). During re-annealing, the N7 part of the second adaptor may hybridize to a segment at a random distance from the first hybridization sequence motif, thereby forming a single stranded loop 1806. In some embodiments, the N7 end of the second adaptor may not hybridize until denaturation produces long single stranded regions of the nucleic acid 1807. The average distance between two captured genomic segments (which are generally from about 20 to about 200 bases in length) will in many embodiments be between about 0.5 to about 20 kilobases. This average distance will depend in part on

the number of degenerate bases ("Ns") of the adaptors and the stringency of hybridization conditions. The re-annealing step can then be followed by another round of adaptor hybridization and nick translation ligation. A final adaptor (in Figure 18, this final adaptor is pictured as a third adaptor 1808, but it will be appreciated that the final adaptor may be the fourth, fifth, sixth, seventh or more adaptor inserted according to any of the methods described herein) is similar to the second adaptor but will in many embodiments lack the degenerate bases at the 3' end. In further embodiments, the final adaptor may comprise a binding site for a primer for an amplification reaction, for example a PCR primer.

[0177] In still further embodiments, amplification reactions, such as PCR reactions (see 1809 in Figure 18), can be carried out, for example, by using primer binding sites included in the first and final adaptors. In still further embodiments, the first and final adaptors may be two arms of the same adaptor and more than one adaptor may be inserted prior to the addition of the final adaptor. In a yet further embodiment, the amplification products may be used to form circular double stranded nucleic acid molecules for further adaptor insertion using any of the process described herein or known in the art.

IVD(iii). Controlled insertion of subsequent adaptors: protection of restriction endonuclease recognition sites

[0178] In addition to controlling the orientation of adaptors inserted into a target nucleic acid as described above, multiple adaptors can also be inserted into a target nucleic acid at specified locations relative to previously inserted adaptors. Such methods include embodiments in which certain restriction endonuclease recognition sites, particularly recognition sites contained in a previously inserted adaptor, are protected from inactivation. In order to ligate subsequent adaptors in a desired position and orientation, the present invention provides methods in which a Type IIs restriction endonuclease binds to a recognition site within the first adaptor of a circular nucleic acid construct and then cleaves at a point outside the first adaptor and in the genomic fragment (also referred to herein as the "target nucleic acid"). A second adaptor can then be ligated into the point at which cleavage occurs (again, usually by adding two adaptor arms of the second adaptor). In order to cleave the target nucleic acid at a known point, it is necessary to block any other recognition sites for that same enzyme that may randomly be encompassed in the target nucleic acid, such that the only point at which that restriction endonuclease can bind is within the first adaptor, thus avoiding undesired cleavage of the constructs. Generally, the recognition site in the first adaptor is first protected from inactivation, and then any other unprotected recognition sites in the construct are inactivated, generally through methylation. By "inactivation" of a restriction endonuclease recognition site herein is meant that the recognition site is somehow rendered unavailable for binding by a restriction endonuclease, thus preventing the downstream step of cleavage by that enzyme. For example, methylated recognition sites will not bind the restriction endonuclease, and thus no cleavage will occur. Once all recognition sites in a nucleic acid construct that are unprotected have been methylated, only the unmethylated recognition site within the adaptor will allow binding of the enzyme with subsequent cleaving. Other methods of inactivating recognition sites include without limitation applying a methylase block to the recognition site, using a blocking oligonucleotide to block the recognition site, using some other blocking molecule, such as a zinc finger protein, to block the recognition site, and nicking the recognition site to prevent methylation. Such methods for protecting the desired recognition site are described in U.S. Application Nos. 12/265,593, filed November 5, 2008 and

12/266,385, filed November 6, 2008, which are both herein incorporated by reference in their entirety and for all purposes and in particular for all teachings related to inserting multiple interspersed adaptors into a target nucleic acid.

[0179] It will be appreciated that the methods described above for controlling the orientation in which adaptors and target nucleic acids ligate to each other may also be used in combination with the methods described below for controlling the spacing of each subsequently added adaptor.

[0180] In one aspect, the present invention provides a method of protecting the recognition site in the first adaptor from inactivation by rendering the recognition site in the first adaptor single-stranded, such that a methylase that is only able to methylate double-stranded molecules will be unable to methylate the recognition site being protected. One method of rendering the recognition site in the first adaptor single-stranded is by amplifying the linear genomic fragments ligated to the two first adaptor arms using primers modified with uracil. The primers are complementary to the adaptor arms and are modified with uracil such that, upon amplification (generally using PCR), the resultant linear constructs contain uracil embedded in the recognition site of one of the first adaptor arms. The primers generate a PCR product with uracils close to the Type IIs restriction endonuclease recognition site in the first and/or second arms of the first adaptor. Digestion of the uracil renders the region(s) of the adaptor arm that include the Type IIs recognition site to be protected single stranded. A sequence specific methylase is then applied to the linear constructs that will methylate all of the double-stranded recognition sites for the same endonuclease as that contained in the first adaptor. Such a sequence-specific methylase will not be able to methylate the single stranded recognition site in the first adaptor arm(s), and thus the recognition site in the first adaptor arm(s) will be protected from inactivation by methylation.

[0181] In some cases, as more fully described below, a single adaptor may have two of the same recognition sites, to allow cleavage both “upstream” and “downstream” from the same adaptor. In this embodiment, as depicted in Figure 7, the primers and uracil positions are chosen appropriately, such that either the “upstream” or “downstream” recognition site may be selectively protected from inactivation or inactivated.

[0182] A third adaptor can be inserted on the other side of the first adaptor by cutting with a restriction endonuclease bound to a recognition site in the second arm of the first adaptor (the recognition site that was originally inactivated by methylation). In order to make this recognition site available, uracil-modified primers complementary to the second recognition site in the first adaptor are used to amplify the circular constructs to produce third linear constructs in which the first adaptor comprises uracils embedded in the second restriction recognition site. The uracils are degraded to render the first adaptor single stranded, which protects the recognition site in the adaptor from methylation. Applying a sequence-specific methylase will then inactivate all unprotected recognition sites. Upon circularization the recognition site in the first adaptor is reconstituted, and applying the restriction endonuclease will cleave the circle, producing a position at which the third adaptor can be inserted in a third linear construct. Ligating third adaptor arms to the third linear construct will follow the same general procedure described above – the third linear construct will be A- or G-tailed, the third adaptor arms will be T- or C-tailed, allowing the adaptor arms to anneal to the third linear construct and be ligated. The linear construct comprising the third adaptor arms is then circularized to form a third circular construct. Like the second adaptor, the third adaptor will generally comprise a recognition site for a restriction endonuclease that is different than the recognition site contained in the first adaptor.

[0183] A fourth adaptor can be added by utilizing Type IIs restriction endonucleases that have recognition sites in the second and third adaptors. Cleavage with these restriction endonucleases will result in a fourth linear construct that can then be ligated to fourth adaptor arms. Circularization of the fourth linear construct ligated to the fourth adaptor arms will produce the nucleic acid template constructs of the invention.

[0184] In general, methods of the invention provide a way to specifically protect a Type IIs endonuclease recognition site from inactivation such that, once all remaining unprotected recognition sites in a construct are inactivated, application of the Type IIs endonuclease will result in binding only to the protected site, thus providing control over where the subsequent cleavage occurs in the construct. The method described above provides one embodiment of how to protect the desired recognition site from inactivation. It will be appreciated that the above-described method can be modified using techniques known in the art, and that such modified methods are encompassed by the present invention.

[0185] In one exemplary embodiment, each subsequently inserted adaptor is inserted using a method in which a recognition site is protected from inactivation using a combination of methods. Figure 19 is a schematic illustration of an embodiment in which a second adaptor is inserted at a desired position relative to a first adaptor by employing a process that is a combination of methylation and protection from methylation using a combination of uracil degradation and nickase. Figure 19 shows genomic DNA of interest 1902 having a Type IIs restriction endonuclease recognition site at 1904. The genomic DNA is fractionated or fragmented in process 1905 to produce fragment 1906 having a Type IIs restriction endonuclease recognition site 1904. Adaptor arms 1908 and 1910 are ligated to fragment 1906 in process 1907. Fragment 1906 with first and second adaptor arms 1908 and 1910 (a library construct) are amplified by PCR in process 1911, using uracil-modified primers 1912 complementary to adaptor arms 1908 and 1910. The primers generate a PCR product with uracils close to the Type IIs restriction endonuclease recognition site. In process 1913, the uracils are specifically degraded using, *e.g.*, uracil-DNA glycosylase enzyme (Krokan, et al., (1997) *Biochem. J.* 325:1-16), leaving a PCR product that is single-stranded in the Type IIs restriction endonuclease recognition site region. As shown, uracil incorporation and degradation may be used to render the Type IIs restriction endonuclease recognition site single-stranded; however, as described further herein, other methods may be employed to render these regions single-stranded including use of 3' or 5' exonucleases in a limited digest.

[0186] In process 1915, a sequence-specific nickase is used to nick bases in each double-stranded Type IIs restriction endonuclease recognition site to protect these sites from Type IIs restriction endonuclease recognition. However, the single-stranded Type IIs restriction endonuclease recognition site portions in first and second adaptor arms 1908 and 1910 will not be nicked, and, once circularized and ligated 1917, the Type IIs restriction endonuclease recognition site in the first and second adaptor arms re-forms such that this Type IIs restriction endonuclease recognition site is available for restriction. When selecting the nickase and the Type IIs restriction endonucleases for this process, it is preferred that the two enzymes recognize the same sequence or that one enzyme recognizes a subsequence (sequence within the sequence) of the other enzyme. Alternatively, the nickase may recognize a different sequence, but is positioned within the adaptor so that it nicks in the Type IIs restriction endonuclease recognition site. Use of uracil or 3' or 5' degradation permits the use of one nickase enzyme throughout the process; alternatively, more than one sequence-specific nickase may be employed. The circularized construct is then cut with the Type IIs restriction endonuclease in process

1919 where the Type IIs restriction endonuclease recognition site is indicated at 1922, the construct is cut at 1920, and the nick is indicated at 1918, resulting in a linearized construct available for ligation of a second set of adaptor arms to be added to the construct in process 1921

[0187] Ligation process 1921 adds first 1924 and second 1926 adaptor arms of the second adaptor to the linearized construct, and a second amplification is performed by PCR at process 1923, again using uracil-modified primers 1928 complementary to adaptor arms 1924 and 1926. As before, the primers generate a PCR product with uracils close to the Type IIs restriction endonuclease recognition site. In process 1925, the uracils are specifically degraded leaving a PCR product that is single-stranded in the Type IIs restriction endonuclease recognition site region of the first and second adaptor arms 1924 and 1926 of the second adaptor. Ligation process 1921 also serves to repair the nick 1918 in the Type IIs restriction site 1904 in the target nucleic acid fragment 1906. In process 1927, the sequence-specific nickase again is used to nick bases in the double-stranded Type IIs restriction endonuclease recognition sites in the target nucleic acid fragment (there is nicking 1914 of the Type IIs restriction endonuclease recognition site 1904) and in the Type IIs restriction endonuclease recognition site of the first adaptor 1930 protecting these sites from Type IIs restriction endonuclease recognition.

[0188] The nicked construct is then circularized and ligated at process 1929, where the Type IIs restriction endonuclease recognition site in the first and second arms 1924 and 1926 of the second adaptor is re-formed 1932 and the process is repeated where the circularized construct is cut again with the Type IIs restriction endonuclease in process 1931 to generate another linearized construct (this one with first and second adaptors already added) available for ligation of a third pair of adaptor arms 1936 and 1938 to the construct. The Type IIs restriction endonuclease recognition site is shown at 1922, the site of restriction is shown at 1920, the nick Type IIs restriction endonuclease recognition site in the target nucleic acid fragment is shown at 1918 and the nick in the first adaptor is shown at 1934. The process can be repeated to add as many adaptors as are desired. As shown here, the first added adaptor had one Type IIs restriction endonuclease recognition site; however, in other aspects, the first added adaptor may have two Type IIs restriction endonuclease recognition sites to allow for precise selection of target nucleic acid size for the construct.

[0189] In one aspect, adaptors can be designed to have sequence-specific nickase sites surrounding or partially overlapping the Type IIs restriction endonuclease recognition site. By utilizing the nickase, the Type IIs restriction endonuclease recognition site(s) of each adaptor can be selectively protected from methylation. In further embodiments, the nickase may recognize another sequence or site, but will cut at the Type IIs restriction endonuclease recognition site. Nickases are endonucleases recognize a specific recognition sequence in double-stranded DNA, and cut one strand at a specific location relative to the recognition sequence, thereby giving rise to single-stranded breaks in duplex DNA and include but are not limited to Nb.BsrDI, Nb.BsmI, Nt.BbvCI, Nb.Bbv.Nb.BtsI and Nt.BstNBI. By employing a combination of sequence-specific nickase and Type IIs restriction endonuclease, all Type IIs restriction endonuclease recognition sites in the target nucleic acid as well as the Type IIs restriction endonuclease recognition sites in any previously-inserted adaptor can be protected from digestion (assuming, of course, the Type IIs restriction endonuclease is nick sensitive, i.e., will not bind at a recognition site that has been nicked).

[0190] Figure 20 is a schematic representation of an embodiment of methods of the invention where a desired position of a second adaptor relative to a first adaptor is selected using methylation and sequence-specific nickases. Figure 20 shows genomic DNA of interest (target nucleic acid) 2002 having

a Type IIs restriction endonuclease recognition site at 2004. The genomic DNA is fractionated or fragmented in process 2005 to produce fragments 2006 having a Type IIs restriction endonuclease recognition site 2004. Adaptor arms 2008 and 2010 are ligated to fragment 2006 in process 2007. Fragment 2006 with adaptor arms 2008 and 2010 (a library construct) is circularized in process 2009 and amplified by circle dependent amplification in process 2011, resulting in a highly-branched concatemer of alternating target nucleic acid fragments 2006 (with the Type IIs restriction endonuclease recognition site at 2004) and first adaptors 2012.

[0191] In process 2013, a sequence-specific nickase 2030 is used to nick the nucleic acid in or near specific Type IIs restriction endonuclease recognition sites in the adaptor in the library construct thereby blocking methylation of these sites. Here, the Type IIs restriction endonuclease recognition sites in adaptor arms 2012 and 2014 are nicked by sequence-specific nickase 2030. In process 2015, un-nicked Type IIs restriction endonuclease recognition sites in the construct are methylated—here, methylation 2016 of the Type IIs restriction endonuclease recognition site 2004)--protecting these sites from Type IIs restriction endonuclease recognition. However, the Type IIs restriction endonuclease recognition sites in adaptors 2012 and 2014 are not methylated due to the presence of the nicks.

[0192] At process 2017, the nicks are repaired in the library construct, resulting in a library construct where the Type IIs restriction endonuclease recognition site in adaptors 2012 are available for recognition and restriction 2018, and the Type IIs restriction endonuclease recognition site in the genomic fragment 2004, is not. The methylated construct is then ligated to an second pair of adaptor arms, circularized, and amplified via circle dependent amplification at process 2021, resulting in a concatemer of alternating target nucleic acid fragments 2006 (with the Type IIs restriction endonuclease recognition site at 2004), first adaptors 2012 and second adaptors 2020. Next, in process 2023, sequence-specific nicking is performed again, this time with a sequence-specific nickase that recognizes a site in the second adaptor 2020 to block methylation of the Type IIs restriction endonuclease recognition site in the second adaptor 2020, but not the other Type IIs restriction endonuclease recognition sites in the construct (i.e., the Type IIs restriction endonuclease recognition site 2004 in the fragment and the Type IIs restriction endonuclease recognition site in first adaptor 2012). The process then continues with methylation 2015, and further adaptor arms are added, if desired. Different sequence-specific nickase sites are used in each different adaptor, allowing for sequence-specific nicking throughout the process.

[0193] Figure 21 is a schematic representation of a process where a desired position of a second adaptor relative to a first adaptor is selected using methylation and sequence-specific methylase blockers. Figure 21 shows genomic DNA of interest (target nucleic acid) 2212 having a Type IIs restriction endonuclease recognition site at 2214. The genomic DNA is fractionated or fragmented in process 2105 to produce fragment 2106 having a Type IIs restriction endonuclease recognition site 2104. Adaptor arms 2108 and 2110 are ligated to fragment 2106 in process 2107. Fragment 2106 with adaptor arms 2108 and 2110 (a library construct) is circularized in process 2109 and amplified by circle dependent amplification in process 2111, resulting in a highly-branched concatemer of alternating target nucleic acid fragments 2106 (with the Type IIs restriction endonuclease recognition site at 2104) and first adaptors 2112.

[0194] In process 2113, a sequence-specific methylase blocker 2130 such as a zinc finger is used to block methylation in specific Type IIs restriction endonuclease recognition sites in the library construct.

Here, the Type IIs restriction endonuclease recognition sites in adaptor arms 2112 and 2114 are blocked by methylase blocker 2130. When selecting the methylase blocker and the Type IIs restriction endonucleases for this process, it is not necessary that the two entities recognize the same site sequence or that one entity recognizes a subsequence of the other entity. The blocker sequences may be up- or downstream from the Type IIs restriction endonuclease recognition site, but are of a configuration that the methylase blocker blocks the site (such as with a zinc finger or other nucleic acid binding protein or other entity). In process 2115, unprotected Type IIs restriction endonuclease recognition sites in the construct are methylated—here, methylation 2116 of the Type IIs restriction endonuclease recognition site 2104)--protecting these sites from Type IIs restriction endonuclease recognition. However, the Type IIs restriction endonuclease recognition sites in adaptors 2112 and 2114 are not methylated due to the presence of the methylase blocker.

[0195] At process 2117, the methylase blocker is released from the library construct, resulting in a library construct where the Type IIs restriction endonuclease recognition site in adaptors 2112 are available for recognition and restriction 2118, and the Type IIs restriction endonuclease recognition site in the genomic fragment 2104, is not. The methylated construct is then ligated to an second pair of adaptor arms, circularized, and amplified via circle dependent amplification at process 2121, resulting in a concatemer of alternating target nucleic acid fragments 2106 (with the Type IIs restriction endonuclease recognition site at 2104), first adaptors 2112 and second adaptors 2120. Next, in process 2123, methylase blocking is performed again, this time with a methylase blocker that recognizes a site in the second adaptor 2120 to block methylation of the Type IIs restriction endonuclease recognition site in the second adaptor 2120, but not the other Type IIs restriction endonuclease recognition sites in the construct (i.e., the Type IIs restriction endonuclease recognition site 2104 in the fragment and the Type IIs restriction endonuclease recognition site in first adaptor 2112). The process then continues with methylation 2115, and further adaptor arms are added, if desired. Different methylase blocker sites are used in each different adaptor, allowing for sequence-specific methylase blocking throughout the process. Though Figures 9 and 21 show insertion of a second adaptor in relation to a first, it should be understood that the process is applicable to adaptors added subsequently to the second adaptor, creating library constructs with up to four, six, eight, ten or more inserted adaptors.

[0196] Figure 22 is a schematic illustration of a process where a desired position of a second adaptor relative to a first adaptor is selected using methylation and uracil degradation. Figure 22 shows genomic DNA of interest 2202 having a Type IIs restriction endonuclease recognition site at 2204. The genomic DNA is fractionated or fragmented in process 2205 to produce fragments 2206 having a Type IIs restriction endonuclease recognition site 2204. Adaptor arms 2208 and 2210 are ligated to fragment 2206 in process 2207. Fragment 2206 with first and second adaptor arms 2208 and 2210 (a library construct) are amplified by PCR in process 2211, using uracil-modified primers 2212 complementary to adaptor arms 2208 and 2210. The primers generate a PCR product with uracils at or close to the Type IIs restriction endonuclease recognition site. In process 2213, the uracils are specifically degraded using, *e.g.*, uracil-DNA glycosylase enzyme (Krokan, et al., (1997) *Biochem. J.* 325:1-16), leaving a PCR product that is single-stranded in the Type IIs restriction endonuclease recognition site region. As shown, uracil incorporation and degradation may be used to render the Type IIs restriction endonuclease recognition site single-stranded; however, as described further herein, other methods may be employed to render these regions single-stranded including use of 3' or 5' exonucleases in a limited digest.

[0197] In process 2215, a sequence-specific methylase is used to methylate bases in each double-stranded Type IIs restriction endonuclease recognition site (here, there is methylation 2214 of the Type IIs restriction endonuclease recognition site 2204), to protect these sites from Type IIs restriction endonuclease recognition. However, the single-stranded Type IIs restriction endonuclease recognition sites in first and second adaptor arms 2208 and 2210 are not methylated, and, once circularized and ligated 2217, the Type IIs restriction endonuclease recognition site re-forms 2216 such that this Type IIs restriction endonuclease recognition site is available for restriction. When selecting the methylase and the Type IIs restriction endonucleases for this process, it is necessary that the two enzymes recognize the same sequence or that one enzyme recognizes a subsequence (sequence within the sequence) of the other enzyme. The circularized construct is then cut with the Type IIs restriction endonuclease in process 2219 where the Type IIs restriction endonuclease recognition site is indicated at 2218 and the construct is cut at 2220, resulting in a linearized construct available for ligation of a second set of adaptor arms to be added to the construct in process 2221.

[0198] Ligation process 2221 adds first 2222 and second 2224 adaptor arms of the second adaptor to the linearized construct, and a second amplification is performed by PCR at process 2223, again using uracil-modified primers 2226 complementary to adaptor arms 2222 and 2224. As before, the primers generate a PCR product with uracils close to the Type IIs restriction endonuclease recognition site. In process 2225, the uracils are specifically degraded leaving a PCR product that is single-stranded in the Type IIs restriction endonuclease recognition site region of the first and second adaptor arms 2222 and 2224 of the second adaptor. In process 2227, the sequence-specific methylase again is used to methylate bases in the double-stranded Type IIs restriction endonuclease recognition sites in the target nucleic acid fragment (again, there is methylation 2214 of the Type IIs restriction endonuclease recognition site 2204) and in the Type IIs restriction endonuclease recognition site of the first adaptor 2228 protecting these sites from Type IIs restriction endonuclease recognition. The methylated construct is then circularized at process 2229, where the Type IIs restriction endonuclease recognition site in the first and second arms 2222 and 2224 of the second adaptor is re-formed 2230 and the process is repeated where the circularized construct is cut again with the Type IIs restriction endonuclease in process 2219 to generate another linearized construct (this one with first and second adaptors already added) available for ligation of a third pair of adaptor arms to the construct. The process can be repeated to add as many adaptors as are desired. As shown here, the first added adaptor had one Type IIs restriction endonuclease recognition site; however, in other aspects, the first added adaptor may have two Type IIs restriction endonuclease recognition sites to allow for precise selection of target nucleic acid size for the construct.

[0199] In addition to the above methods for controlled insertion of multiple interspersed adaptors, constructs comprising adaptors in specific orientations may further be selected by enriching a population of constructs for those with adaptors in the desired orientations. Such enrichment methods are described in US Ser. Nos. 60/864,992 filed 11/09/06; 11/943,703, filed 11/02/07; 11/943,697, filed 11/02/07; 11/943,695, filed 11/02/07; and PCT/US07/835540; filed 11/02/07, all of which are incorporated by reference in their entirety for all purposes and in particular for all teachings related to methods and compositions for selecting for specific orientations of adaptors.

V. Making DNBs

[0200] In one aspect, nucleic acid templates of the invention are used to generate nucleic acid nanoballs, which are also referred to herein as "DNA nanoballs," "DNBs", and "amplicons". These nucleic acid nanoballs are generally concatemers comprising multiple copies of a nucleic acid template of the invention, although nucleic acid nanoballs of the invention may be formed from any nucleic acid molecule using the methods described herein.

[0201] In one aspect, rolling circle replication (RCR) is used to create concatemers of the invention. The RCR process has been shown to generate multiple continuous copies of the M13 genome. (Blanco, et al., (1989) *J Biol Chem* 264:8935-8940). In such a method, a nucleic acid is replicated by linear concatemerization. Guidance for selecting conditions and reagents for RCR reactions is available in many references available to those of ordinary skill, including U.S. Patent Nos. 5,426,180; 5,854,033; 6,143,495; and 5,871,921, each of which is hereby incorporated by reference in its entirety for all purposes and in particular for all teachings related to generating concatemers using RCR or other methods.

[0202] Generally, RCR reaction components include single stranded DNA circles, one or more primers that anneal to DNA circles, a DNA polymerase having strand displacement activity to extend the 3' ends of primers annealed to DNA circles, nucleoside triphosphates, and a conventional polymerase reaction buffer. Such components are combined under conditions that permit primers to anneal to DNA circle. Extension of these primers by the DNA polymerase forms concatemers of DNA circle complements. In some embodiments, nucleic acid templates of the invention are double stranded circles that are denatured to form single stranded circles that can be used in RCR reactions.

[0203] In some embodiments, amplification of circular nucleic acids may be implemented by successive ligation of short oligonucleotides, *e.g.,* 6-mers, from a mixture containing all possible sequences, or if circles are synthetic, a limited mixture of these short oligonucleotides having selected sequences for circle replication, a process known as "circle dependent amplification" (CDA). "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 being formed.

[0204] Concatemers may also be generated by ligation of target DNA in the presence of a bridging template DNA complementary to both beginning and end of the target molecule. A population of different target DNA may be converted in concatemers by a mixture of corresponding bridging templates.

[0205] In some embodiments, a subset of a population of nucleic acid templates may be isolated based on a particular feature, such as a desired number or type of adaptor. This population can be isolated or otherwise processed (*e.g.,* size selected) using conventional techniques, *e.g.,* a conventional spin column, or the like, to form a population from which a population of concatemers can be created using techniques such as RCR.

[0206] Methods for forming DNBs of the invention are described in Published Patent Application Nos. WO2007120208, WO2006073504, WO2007133831, and US2007099208, and U.S. Patent Application Nos. 60/992,485; 61/026,337; 61/035,914; 61/061,134; 61/116,193; 61/102,586; 12/265,593; 12/266,385; 11/938,096; 11/981,804; No 11/981,797; 11/981,793; 11/981,767; 11/981,761; 11/981,730, filed October 31, 2007; 11/981,685; 11/981,661; 11/981,607; 11/981,605; 11/927,388; 11/927,356; 11/679,124; 11/541,225; 10/547,214; 11/451,692; and 11/451,691, all of which are incorporated herein by reference in their entirety for all purposes and in particular for all teachings related to forming DNBs.

VI. Producing arrays of DNBs

[0207] In one aspect, DNBs of the invention are disposed on a surface to form a random array of single molecules. DNBs can be fixed to surface by a variety of techniques, including covalent attachment and non-covalent attachment. In one embodiment, a surface may include capture probes that form complexes, *e.g.,* double stranded duplexes, with component of a polynucleotide molecule, such as an adaptor oligonucleotide. In other embodiments, capture probes may comprise oligonucleotide clamps, or like structures, that form triplexes with adaptors, as described in Gryaznov et al, U.S. patent 5,473,060, which is hereby incorporated in its entirety.

[0208] Methods for forming arrays of DNBs of the invention are described in Published Patent Application Nos. WO2007120208, WO2006073504, WO2007133831, and US2007099208, and U.S. Patent Application Nos. 60/992,485; 61/026,337; 61/035,914; 61/061,134; 61/116,193; 61/102,586; 12/265,593; 12/266,385; 11/938,096; 11/981,804; No 11/981,797; 11/981,793; 11/981,767; 11/981,761; 11/981,730; 11/981,685; 11/981,661; 11/981,607; 11/981,605; 11/927,388; 11/927,356; 11/679,124; 11/541,225; 10/547,214; 11/451,692; and 11/451,691, all of which are incorporated herein by reference in their entirety for all purposes and in particular for all teachings related to forming arrays of DNBs.

[0209] In some embodiments, a surface may have reactive functionalities that react with complementary functionalities on the polynucleotide molecules to form a covalent linkage, *e.g.,* by way of the same techniques used to attach cDNAs to microarrays, *e.g.,* Smirnov et al (2004), Genes, Chromosomes & Cancer, 40: 72-77; Beaucage (2001), Current Medicinal Chemistry, 8: 1213-1244, which are incorporated herein by reference. DNBs may also be efficiently attached to hydrophobic surfaces, such as a clean glass surface that has a low concentration of various reactive functionalities, such as –OH groups. Attachment through covalent bonds formed between the polynucleotide molecules and reactive functionalities on the surface is also referred to herein as “chemical attachment”.

[0210] In still further embodiments, polynucleotide molecules can adsorb to a surface. In such an embodiment, the polynucleotide molecules are immobilized through non-specific interactions with the surface, or through non-covalent interactions such as hydrogen bonding, van der Waals forces, and the like.

[0211] Attachment may also include wash steps of varying stringencies to remove incompletely attached single molecules or other reagents present from earlier preparation steps whose presence is undesirable or that are nonspecifically bound to surface.

[0212] In one aspect, DNBs on a surface are confined to an area of a discrete region. Discrete regions may be incorporated into a surface using methods known in the art and described further herein. In exemplary embodiments, discrete regions contain reactive functionalities or capture probes which can be used to immobilize the polynucleotide molecules.

[0213] The discrete regions may have defined locations in a regular array, which may correspond to a rectilinear pattern, hexagonal pattern, or the like. A regular array of such regions is advantageous for detection and data analysis of signals collected from the arrays during an analysis. Also, first- and/or second-stage amplicons confined to the restricted area of a discrete region provide a more concentrated or intense signal, particularly when fluorescent probes are used in analytical operations, thereby providing higher signal-to-noise values. In some embodiments, DNBs are randomly distributed on the discrete regions so that a given region is equally likely to receive any of the different single molecules. In other words, the resulting arrays are not spatially addressable immediately upon fabrication, but may be made so by carrying out an identification, sequencing and/or decoding operation. As such, the identities of the polynucleotide molecules of the invention disposed on a surface are discernable, but not initially known upon their disposition on the surface. In some embodiments, the area of discrete is selected, along with attachment chemistries, macromolecular structures employed, and the like, to correspond to the size of single molecules of the invention so that when single molecules are applied to surface substantially every region is occupied by no more than one single molecule. In some embodiments, DNBs are disposed on a surface comprising discrete regions in a patterned manner, such that specific DNBs (identified, in an exemplary embodiment, by tag adaptors or other labels) are disposed on specific discrete regions or groups of discrete regions.

[0214] In some embodiments, the area of discrete regions is less than 1 μm^2; and in some embodiments, the area of discrete regions is in the range of from 0.04 μm^2 to 1 μm^2; and in some embodiments, the area of discrete regions is in the range of from 0.2 μm^2 to 1 μm^2. In embodiments in which discrete regions are approximately circular or square in shape so that their sizes can be indicated by a single linear dimension, the size of such regions are in the range of from 125 nm to 250 nm, or in the range of from 200 nm to 500 nm. In some embodiments, center-to-center distances of nearest neighbors of discrete regions are in the range of from 0.25 μm to 20 μm; and in some embodiments, such distances are in the range of from 1 μm to 10 μm, or in the range from 50 to 1000 nm. Generally, discrete regions are designed such that a majority of the discrete regions on a surface are optically resolvable. In some embodiments, regions may be arranged on a surface in virtually any pattern in which regions have defined locations.

[0215] In further embodiments, molecules are directed to the discrete regions of a surface, because the areas between the discrete regions, referred to herein as "inter-regional areas," are inert, in the sense that concatemers, or other macromolecular structures, do not bind to such regions. In some embodiments, such inter-regional areas may be treated with blocking agents, *e.g.,* DNAs unrelated to concatemer DNA, other polymers, and the like.

[0216] A wide variety of supports may be used with the compositions and methods of the invention to form random arrays. In one aspect, supports are rigid solids that have a surface, preferably a substantially planar surface so that single molecules to be interrogated are in the same plane. The latter feature permits efficient signal collection by detection optics, for example. In another aspect, the support comprises beads, wherein the surface of the beads comprise reactive functionalities or capture probes that can be used to immobilize polynucleotide molecules.

[0217] In still another aspect, solid supports of the invention are nonporous, particularly when random arrays of single molecules are analyzed by hybridization reactions requiring small volumes. Suitable

solid support materials include materials such as glass, polyacrylamide-coated glass, ceramics, silica, silicon, quartz, various plastics, and the like. In one aspect, the area of a planar surface may be in the range of from 0.5 to 4 cm^2. In one aspect, the solid support is glass or quartz , such as a microscope slide, having a surface that is uniformly silanized. This may be accomplished using conventional protocols, *e.g.,* acid treatment followed by immersion in a solution of 3-glycidoxypropyl trimethoxysilane, N,N-diisopropylethylamine, and anhydrous xylene (8:1:24 v/v) at 80°C, which forms an epoxysilanized surface. *e.g.,* Beattie et a (1995), Molecular Biotechnology, 4: 213. Such a surface is readily treated to permit end-attachment of capture oligonucleotides, *e.g.,* by providing capture oligonucleotides with a 3' or 5' triethylene glycol phosphoryl spacer (see Beattie et al, cited above) prior to application to the surface. Further embodiments for functionalizing and further preparing surfaces for use in the present invention are described for example in U.S. Patent Application Ser. Nos. 60/992,485; 61/026,337; 61/035,914; 61/061,134; 61/116,193; 61/102,586; 12/265,593; 12/266,385; 11/938,096; 11/981,804; No 11/981,797; 11/981,793; 11/981,767; 11/981,761; 11/981,730; 11/981,685; 11/981,661; 11/981,607; 11/981,605; 11/927,388; 11/927,356; 11/679,124; 11/541,225; 10/547,214; 11/451,692; and 11/451,691, each of which is herein incorporated by reference in its entirety for all purposes and in particular for all teachings related to preparing surfaces for forming arrays and for all teachings related to forming arrays, particularly arrays of DNBs.

[0218] In embodiments of the invention in which patterns of discrete regions are required, photolithography, electron beam lithography, nano imprint lithography, and nano printing may be used to generate such patterns on a wide variety of surfaces, *e.g.,* Pirrung et al, U.S. patent 5,143,854; Fodor et al, U.S. patent 5,774,305; Guo, (2004) Journal of Physics D: Applied Physics, 37: R123-141; which are incorporated herein by reference.

[0219] In one aspect, surfaces containing a plurality of discrete regions are fabricated by photolithography. A commercially available, optically flat, quartz substrate is spin coated with a 100-500 nm thick layer of photo-resist. The photo-resist is then baked on to the quartz substrate. An image of a reticle with a pattern of regions to be activated is projected onto the surface of the photo-resist, using a stepper. After exposure, the photo-resist is developed, removing the areas of the projected pattern which were exposed to the UV source. This is accomplished by plasma etching, a dry developing technique capable of producing very fine detail. The substrate is then baked to strengthen the remaining photo-resist. After baking, the quartz wafer is ready for functionalization. The wafer is then subjected to vapor-deposition of 3-aminopropyldimethylethoxysilane. The density of the amino functionalized monomer can be tightly controlled by varying the concentration of the monomer and the time of exposure of the substrate. Only areas of quartz exposed by the plasma etching process may react with and capture the monomer. The substrate is then baked again to cure the monolayer of amino-functionalized monomer to the exposed quartz. After baking, the remaining photo-resist may be removed using acetone. Because of the difference in attachment chemistry between the resist and silane, aminosilane-functionalized areas on the substrate may remain intact through the acetone rinse. These areas can be further functionalized by reacting them with p-phenylenediisothiocyanate in a solution of pyridine and N-N-dimethlyformamide. The substrate is then capable of reacting with amine-modified oligonucleotides. Alternatively, oligonucleotides can be prepared with a 5'-carboxy-modifier-c10 linker (Glen Research). This technique allows the oligonucleotide to be attached directly to the amine modified support, thereby avoiding additional functionalization steps.

[0220] In another aspect, surfaces containing a plurality of discrete regions are fabricated by nano-imprint lithography (NIL). For DNA array production, a quartz substrate is spin coated with a layer of resist, commonly called the transfer layer. A second type of resist is then applied over the transfer layer, commonly called the imprint layer. The master imprint tool then makes an impression on the imprint layer. The overall thickness of the imprint layer is then reduced by plasma etching until the low areas of the imprint reach the transfer layer. Because the transfer layer is harder to remove than the imprint layer, it remains largely untouched. The imprint and transfer layers are then hardened by heating. The substrate is then put into a plasma etcher until the low areas of the imprint reach the quartz. The substrate is then derivatized by vapor deposition as described above.

[0221] In another aspect, surfaces containing a plurality of discrete regions are fabricated by nano printing. This process uses photo, imprint, or e-beam lithography to create a master mold, which is a negative image of the features required on the print head. Print heads are usually made of a soft, flexible polymer such as polydimethylsiloxane (PDMS). This material, or layers of materials having different properties, are spin coated onto a quartz substrate. The mold is then used to emboss the features onto the top layer of resist material under controlled temperature and pressure conditions. The print head is then subjected to a plasma based etching process to improve the aspect ratio of the print head, and eliminate distortion of the print head due to relaxation over time of the embossed material. Random array substrates are manufactured using nano-printing by depositing a pattern of amine modified oligonucleotides onto a homogenously derivatized surface. These oligonucleotides would serve as capture probes for the RCR products. One potential advantage to nano-printing is the ability to print interleaved patterns of different capture probes onto the random array support. This would be accomplished by successive printing with multiple print heads, each head having a differing pattern, and all patterns fitting together to form the final structured support pattern. Such methods allow for some positional encoding of DNA elements within the random array. For example, control concatemers containing a specific sequence can be bound at regular intervals throughout a random array.

[0222] In still another aspect, a high density array of capture oligonucleotide spots of sub micron size is prepared using a printing head or imprint-master prepared from a bundle, or bundle of bundles, of about 10,000 to 100 million optical fibers with a core and cladding material. By pulling and fusing fibers a unique material is produced that has about 50-1000 nm cores separated by a similar or 2-5 fold smaller or larger size cladding material. By differential etching (dissolving) of cladding material a nano-printing head is obtained having a very large number of nano-sized posts. This printing head may be used for depositing oligonucleotides or other biological (proteins, oligopeptides, DNA, aptamers) or chemical compounds such as silane with various active groups. In one embodiment the glass fiber tool is used as a patterned support to deposit oligonucleotides or other biological or chemical compounds. In this case only posts created by etching may be contacted with material to be deposited. Also, a flat cut of the fused fiber bundle may be used to guide light through cores and allow light-induced chemistry to occur only at the tip surface of the cores, thus eliminating the need for etching. In both cases, the same support may then be used as a light guiding/collection device for imaging fluorescence labels used to tag oligonucleotides or other reactants. This device provides a large field of view with a large numerical aperture (potentially >1). Stamping or printing tools that perform active material or oligonucleotide deposition may be used to print 2 to 100 different oligonucleotides in an interleaved pattern. This process requires precise positioning of the print head to about 50-500 nm. This type of oligonucleotide array may

be used for attaching 2 to 100 different DNA populations such as different source DNA. They also may be used for parallel reading from sub-light resolution spots by using DNA specific anchors or tags. Information can be accessed by DNA specific tags, *e.g.,* 16 specific anchors for 16 DNAs and read 2 bases by a combination of 5-6 colors and using 16 ligation cycles or one ligation cycle and 16 decoding cycles. This way of making arrays is efficient if limited information (*e.g.,* a small number of cycles) is required per fragment, thus providing more information per cycle or more cycles per surface.

[0223] In one aspect, multiple arrays of the invention may be placed on a single surface. For example, patterned array substrates may be produced to match the standard 96 or 384 well plate format. A production format can be an 8 x 12 pattern of 6mm x 6mm arrays at 9mm pitch or 16x24 of 3.33mm x 3.33mm array at 4.5mm pitch, on a single piece of glass or plastic and other optically compatible material. In one example each 6mm x 6mm array consists of 36 million 250-500nm square regions at 1 micrometer pitch. Hydrophobic or other surface or physical barriers may be used to prevent mixing different reactions between unit arrays.

[0224] Other methods of forming arrays of molecules are known in the art and are applicable to forming arrays of DNBs.

[0225] As will be appreciated, a wide range of densities of DNBs and/or nucleic acid templates of the invention can be placed on a surface comprising discrete regions to form an array. In some embodiments, each discrete region may comprise from about 1 to about 1000 molecules. In further embodiments, each discrete region may comprise from about 10 to about 900, about 20 to about 800, about 30 to about 700, about 40 to about 600, about 50 to about 500, about 60 to about 400, about 70 to about 300, about 80 to about 200, and about 90 to about 100 molecules.

[0226] In some embodiments, arrays of nucleic acid templates and/or DNBs are provided in densities of at least 0.5, 1, 2, 3, 4, 5, 6, 7, 8, 9, or 10 million molecules per square millimeter.

VII. Methods of using DNBs

[0227] DNBs made according to the methods described above offer an advantage in identifying sequences in target nucleic acids, because the adaptors contained in the DNBs provide points of known sequence that allow spatial orientation and sequence determination when combined with methods utilizing anchor and sequencing probes. Methods of using DNBs in accordance with the present invention include sequencing and detecting specific sequences in target nucleic acids (e.g., detecting particular target sequences (e.g. specific genes) and/or identifying and/or detecting SNPs). The methods described herein can also be used to detect nucleic acid rearrangements and copy number variation. Nucleic acid quantification, such as digital gene expression (i.e., analysis of an entire transcriptome – all mRNA present in a sample) and detection of the number of specific sequences or groups of sequences in a sample, can also be accomplished using the methods described herein. Although the majority of the discussion herein is directed to identifying sequences of DNBs, it will be appreciated that other, non-concatemeric nucleic acid constructs comprising adaptors may also be used in the embodiments described herein.

VIIA. Overview of cPAL sequencing

[0228] Sequences of DNBs are generally identified in accordance with the present invention using methods referred to herein as combinatorial probe anchor ligation ("cPAL") and variations thereof, as described below. In brief, cPAL involves identifying a nucleotide at a particular detection position in a target nucleic acid by detecting a probe ligation product formed by ligation of at least one anchor probe that hybridizes to all or part of an adaptor and a sequencing probe that contains a particular nucleotide at an "interrogation position" that corresponds to (e.g. will hybridize to) the detection position. The sequencing probe contains a unique identifying label. If the nucleotide at the interrogation position is complementary to the nucleotide at the detection position, ligation can occur, resulting in a ligation product containing the unique label which is then detected. Descriptions of different exemplary embodiments of cPAL methods are provided below. It will be appreciated that the following descriptions are not meant to be limiting and that variations of the following embodiments are encompassed by the present invention.

[0229] "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%.

[0230] As used herein, "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 5x 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. Further examples of stringent conditions are well known in the art, see for example Sambrook J et al. (2001), *Molecular Cloning, A Laboratory Manual*, (3rd Ed., Cold Spring Harbor Laboratory Press.

[0231] As used herein, the term "T_m" generally refers to 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 T_m 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. (2001), *Molecular Cloning, A Laboratory Manual*, (3rd Ed., Cold Spring Harbor Laboratory Press). 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).

[0232] In one example of a cPAL method, referred to herein as "single cPAL", as illustrated in Figure 23, anchor probe 2302 hybridizes to a complementary region on adaptor 2308 of the DNB 2301. Anchor probe 2302 hybridizes to the adaptor region directly adjacent to target nucleic acid 2309, but in some cases, anchor probes can be designed to "reach into" the target nucleic acid by incorporating a desired number of degenerate bases at the terminus of the anchor probe, as is schematically illustrated in Figure 24 and described further below. A pool of differentially labeled sequencing probes 2305 will hybridize to complementary regions of the target nucleic acid, and sequencing probes that hybridize adjacent to anchor probes are ligated to form a probe ligation product, usually by application of a ligase. The sequencing probes are generally sets or pools of oligonucleotides comprising two parts: different nucleotides at the interrogation position, and then all possible bases (or a universal base) at the other positions; thus, each probe represents each base type at a specific position. The sequencing probes are labeled with a detectable label that differentiates each sequencing probe from the sequencing probes with other nucleotides at that position. Thus, in the example illustrated in Figure 23, a sequencing probe 2310 that hybridizes adjacent to anchor probe 2302 and is ligated to the anchor probe will identify the base at a position in the target nucleic acid 5 bases from the adaptor as a "G". Figure 23 depicts a situation where the interrogation base is 5 bases in from the ligation site, but as more fully described below, the interrogation base can also be "closer" to the ligation site, and in some cases at the point of ligation. Once ligated, non-ligated anchor and sequencing probes are washed away, and the presence of the ligation product on the array is detected using the label. Multiple cycles of anchor probe and sequencing probe hybridization and ligation can be used to identify a desired number of bases of the target nucleic acid on each side of each adaptor in a DNB. Hybridization of the anchor probe and the sequencing probe may occur sequentially or simultaneously. The fidelity of the base call relies in part on the fidelity of the ligase, which generally will not ligate if there is a mismatch close to the ligation site.

[0233] The present invention also provides methods in which two or more anchor probes are used in every hybridization-ligation cycle. Figure 25 illustrate an additional example of a "double cPAL with overhang" method in which a first anchor probe 2502 and a second anchor probe 2505 each hybridize to complimentary regions of an adaptor. In the example illustrated in Figure 25, the first anchor probe 2502 is fully complementary to a first region of the adaptor 2511, and the second anchor probe 2505 is complementary to a second adaptor region adjacent to the hybridization position of the first anchor probe. The second anchor probe also comprises degenerate bases at the terminus that is not adjacent to the first anchor probe. As a result, the second anchor probe is able to hybridize to a region of the

target nucleic acid 2512 adjacent to adaptor 2511 (the "overhang" portion). The second anchor probe is generally too short to be maintained alone in its duplex hybridization state, but upon ligation to the first anchor probe it forms a longer anchor probe that is stably hybridized for subsequent methods. As discussed above for the "single cPAL" method, a pool of sequencing probes 2508 that represents each base type at a detection position of the target nucleic acid and labeled with a detectable label that differentiates each sequencing probe from the sequencing probes with other nucleotides at that position is hybridized 2509 to the adaptor-anchor probe duplex and ligated to the terminal 5' or 3' base of the ligated anchor probes. In the example illustrated in Figure 25, the sequencing probes are designed to interrogate the base that is five positions 5' of the ligation point between the sequencing probe 2514 and the ligated anchor probes 2513. Since the second adaptor probe 2505 has five degenerate bases at its 5' end, it reaches five bases into the target nucleic acid 2512, allowing interrogation with the sequencing probe at a full ten bases from the interface between the target nucleic acid 2512 and the adaptor 2511.

[0234] In variations of the above described examples of a double cPAL method, if the first anchor probe terminates closer to the end of the adaptor, the second adaptor probe will be proportionately more degenerate and therefore will have a greater potential to not only ligate to the end of the first adaptor probe but also to ligate to other second adaptor probes at multiple sites on the DNB. To prevent such ligation artifacts, the second anchor probes can be selectively activated to engage in ligation to a first anchor probe or to a sequencing probe. Such activation methods are described in further detail below, and include methods such as selectively modifying the termini of the anchor probes such that they are able to ligate only to a particular anchor probe or sequencing probe in a particular orientation with respect to the adaptor.

[0235] Similar to the double cPAL method described above, it will be appreciated that cPAL methods utilizing three or more anchor probes are also encompassed by the present invention.

[0236] In addition, sequencing reactions can be done at one or both of the termini of each adaptor, e.g., the sequencing reactions can be "unidirectional" with detection occurring 3' or 5' of the adaptor or the other or the reactions can be "bidirectional" in which bases are detected at detection positions 3' and 5' of the adaptor. Bidirectional sequencing reactions can occur simultaneously – i.e., bases on both sides of the adaptor are detected at the same time – or sequentially in any order.

[0237] Multiple cycles of cPAL (whether single, double, triple, etc.) will identify multiple bases in the regions of the target nucleic acid adjacent to the adaptors. In brief, the cPAL methods are repeated for interrogation of multiple adjacent bases within a target nucleic acid by cycling anchor probe hybridization and enzymatic ligation reactions with sequencing probe pools designed to detect nucleotides at varying positions removed from the interface between the adaptor and target nucleic acid. In any given cycle, the sequencing probes used are designed such that the identity of one or more of bases at one or more positions is correlated with the identity of the label attached to that sequencing probe. Once the ligated sequencing probe (and hence the base(s) at the interrogation position(s) is detected, the ligated complex is stripped off of the DNB and a new cycle of adaptor and sequencing probe hybridization and ligation is conducted.

[0238] As will be appreciated, DNBs of the invention can be used in other sequencing methods in addition to the cPAL methods described above, including other sequencing by ligation methods as well as other sequencing methods, including without limitation sequencing by hybridization, sequencing by

synthesis (including sequencing by primer extension), chained sequencing by ligation of cleavable probes, and the like.

[0239] Methods similar to those described above for sequencing can also be used to detect specific sequences in a target nucleic acid, including detection of single nucleotide polymorphisms (SNPs). In such methods, sequencing probes that will hybridize to a particular sequence, such as a sequence containing a SNP, will be applied. Such sequencing probes can be differentially labeled to identify which SNP is present in the target nucleic acid. Anchor probes can also be used in combination with such sequencing probes to provide further stability and specificity.

VIIB. Sequencing

[0240] In one aspect, the present invention provides methods for identifying sequences of DNBs by utilizing sequencing by ligation methods. In one aspect, the present invention provides methods for identifying sequences of DNBs that utilize a combinatorial probe anchor ligation (cPAL) method. Generally, cPAL involves identifying a nucleotide at a detection position in a target nucleic acid by detecting a probe ligation product formed by ligation of an anchor probe and a sequencing probe. Methods of the invention can be used to sequence a portion or the entire sequence of the target nucleic acid contained in a DNB, and many DNBs that represent a portion or all of a genome.

[0241] As discussed further herein, every DNB comprises repeating monomeric units, each monomeric unit comprising one or more adaptors and a target nucleic acid. The target nucleic acid comprises a plurality of detection positions. The term “detection position” refers to a position in a target sequence for which sequence information is desired. As will be appreciated by those in the art, generally a target sequence has multiple detection positions for which sequence information is required, for example in the sequencing of complete genomes as described herein. In some cases, for example in SNP analysis, it may be desirable to just read a single SNP in a particular area.

[0242] The present invention provides methods of sequencing that utilize a combination of anchor probes and sequencing probes. By “sequencing probe” as used herein is meant an oligonucleotide that is designed to provide the identity of a nucleotide at a particular detection position of a target nucleic acid. Sequencing probes hybridize to domains within target sequences, e.g. a first sequencing probe may hybridize to a first target domain, and a second sequencing probe may hybridize to a second target domain. The terms "first target domain" and "second target domain" or grammatical equivalents herein means two portions of a target sequence within a nucleic acid which is under examination. The first target domain may be directly adjacent to the second target domain, or the first and second target domains may be separated by an intervening sequence, for example an adaptor. The terms "first" and "second" are not meant to confer an orientation of the sequences with respect to the 5'-3' orientation of the target sequence. For example, assuming a 5'-3' orientation of the complementary target sequence, the first target domain may be located either 5' to the second domain, or 3' to the second domain. Sequencing probes can overlap, e.g. a first sequencing probe can hybridize to the first 6 bases adjacent to one terminus of an adaptor, and a second sequencing probe can hybrdize to the 4rd-9th bases from the terminus of the adaptor (for example when an anchor probe has three degenerate bases). Alternatively, a first sequencing probe can hybridize to the 6 bases adjacent to the “upstream” terminus

of an adaptor and a second sequencing probe can hybridize to the 6 bases adjacent to the "downstream" terminus of an adaptor.

[0243] Sequencing probes will generally comprise a number of degenerate bases and a specific nucleotide at a specific location within the probe to query the detection position (also referred to herein as an "interrogation position").

[0244] In general, pools of sequencing probes are used when degenerate bases are used. That is, a probe having the sequence "NNN**A**NN" is actually a set of probes of having all possible combinations of the four nucleotide bases at five positions (i.e., 1024 sequences) with an adenosine at the 6th position. (As noted herein, this terminology is also applicable to adaptor probes: for example, when an adaptor probe has "three degenerate bases", for example, it is actually a set of adaptor probes comprising the sequence corresponding to the anchor site, and all possible combinations at 3 positions, so it is a pool of 64 probes).

[0245] In some embodiments, for each interrogation position, four differently labeled pools can be combined in a single pool and used in a sequencing step. Thus, in any particular sequencing step, 4 pools are used, each with a different specific base at the interrogation position and with a different label corresponding to the base at the interrogation position. That is, sequencing probes are also generally labeled such that a particular nucleotide at a particular interrogation position is associated with a label that is different from the labels of sequencing probes with a different nucleotide at the same interrogation position. For example, four pools can be used: NNN**A**NN-dye1, NNN**T**NN-dye2, NNN**C**NN-dye3 and NNN**G**NN-dye4 in a single step, as long as the dyes are optically resolvable. In some embodiments, for example for SNP detection, it may only be necessary to include two pools, as the SNP call will be either a C or an A, etc. Similarly, some SNPs have three possibilities. Alternatively, in some embodiments, if the reactions are done sequentially rather than simultaneously, the same dye can be done, just in different steps: e.g. the NNN**A**NN-dye1 probe can be used alone in a reaction, and either a signal is detected or not, and the probes washed away; then a second pool, NNN**T**NN-dye1 can be introduced.

[0246] In any of the sequencing methods described herein, sequencing probes may have a wide range of lengths, including about 3 to about 25 bases. In further embodiments, sequencing probes may have lengths in the range of about 5 to about 20, about 6 to about 18, about 7 to about 16, about 8 to about 14, about 9 to about 12, and about 10 to about 11 bases.

[0247] Sequencing probes of the present invention are designed to be complementary, and in general, perfectly complementary, to a sequence of the target sequence such that hybridization of a portion target sequence and probes of the present invention occurs. In particular, it is important that the interrogation position base and the detection position base be perfectly complementary and that the methods of the invention do not result in signals unless this is true.

[0248] In many embodiments, sequencing probes are perfectly complementary to the target sequence to which they hybridize; that is, the experiments are run under conditions that favor the formation of perfect basepairing, as is known in the art. As will be appreciated by those in the art, a sequencing probe that is perfectly complementary to a first domain of the target sequence could be only substantially complementary to a second domain of the same target sequence; that is, the present invention relies in many cases on the use of sets of probes, for example, sets of hexamers, that will be perfectly complementary to some target sequences and not to others.

[0249] In some embodiments, depending on the application, the complementarity between the sequencing probe and the target need not be perfect; there may be any number of base pair mismatches, which will interfere with hybridization between the target sequence and the single stranded nucleic acids of the present invention. However, if the number of mismatches is so great that no hybridization can occur under even the least stringent of hybridization conditions, the sequence is not a complementary target sequence. Thus, by "substantially complementary" herein is meant that the sequencing probes are sufficiently complementary to the target sequences to hybridize under normal reaction conditions. However, for most applications, the conditions are set to favor probe hybridization only if perfectly complementarity exists. Alternatively, sufficient complementarity is required to allow the ligase reaction to occur; that is, there may be mismatches in some part of the sequence but the interrogation position base should allow ligation only if perfect complementarity at that position occurs.

[0250] In some cases, in addition to or instead of using degenerate bases in probes of the invention, universal bases which hybridize to more than one base can be used. For example, inosine can be used. Any combination of these systems and probe components can be utilized.

[0251] Sequencing probes of use in methods of the present invention are usually detectably labeled. By "label" or "labeled" herein is meant that a compound has at least one element, isotope or chemical compound attached to enable the detection of the compound. In general, labels of use in the invention include without limitation isotopic labels, which may be radioactive or heavy isotopes, magnetic labels, electrical labels, thermal labels, colored and luminescent dyes, enzymes and magnetic particles as well. Dyes of use in the invention may be chromophores, phosphors or fluorescent dyes, which due to their strong signals provide a good signal-to-noise ratio for decoding. Sequencing probes may also be labeled with quantum dots, fluorescent nanobeads or other constructs that comprise more than one molecule of the same fluorophore. Labels comprising multiple molecules of the same fluorophore will generally provide a stronger signal and will be less sensitive to quenching than labels comprising a single molecule of a fluorophore. It will be understood that any discussion herein of a label comprising a fluorophore will apply to labels comprising single and multiple fluorophore molecules.

[0252] Many embodiments of the invention include the use of fluorescent labels. Suitable dyes for use in the invention include, but are not limited to, fluorescent lanthanide complexes, including those of Europium and Terbium, fluorescein, rhodamine, tetramethylrhodamine, eosin, erythrosin, coumarin, methyl-coumarins, pyrene, Malacite green, stilbene, Lucifer Yellow, Cascade Blue.TM., Texas Red, and others described in the 6th Edition of the Molecular Probes Handbook by Richard P. Haugland, hereby expressly incorporated by reference in its entirety for all purposes and in particular for its teachings regarding labels of use in accordance with the present invention. Commercially available fluorescent dyes for use with any nucleotide for incorporation into nucleic acids include, but are not limited to: Cy3, Cy5, (Amersham Biosciences, Piscataway, New Jersey, USA), fluorescein, tetramethylrhodamine-, Texas Red®, Cascade Blue®, BODIPY® FL-14, BODIPY®R, BODIPY® TR-14, Rhodamine GreenTM, Oregon Green® 488, BODIPY® 630/650, BODIPY® 650/665-, Alexa Fluor® 488, Alexa Fluor® 532, Alexa Fluor® 568, Alexa Fluor® 594, Alexa Fluor® 546 (Molecular Probes, Inc. Eugene, OR, USA), Quasar 570, Quasar 670, Cal Red 610 (BioSearch Technologies, Novato, Ca). Other fluorophores available for post-synthetic attachment include, inter alia, Alexa Fluor® 350, Alexa Fluor® 532, Alexa Fluor® 546, Alexa Fluor® 568, Alexa Fluor® 594, Alexa Fluor® 647, BODIPY 493/503, BODIPY FL, BODIPY R6G, BODIPY 530/550, BODIPY TMR, BODIPY 558/568, BODIPY 558/568, BODIPY 564/570,

BODIPY 576/589, BODIPY 581/591, BODIPY 630/650, BODIPY 650/665, 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 (available from Molecular Probes, Inc., Eugene, OR, USA), and Cy2, Cy3.5, Cy5.5, and Cy7 (Amersham Biosciences, Piscataway, NJ USA, and others). In some embodiments, the labels used include fluoroscein, Cy3, Texas Red, Cy5, Quasar 570, Quasar 670 and Cal Red 610 are used in methods of the present invention.

[0253] Labels can be attached to nucleic acids to form the labeled sequencing probes of the present invention using methods known in the art, and to a variety of locations of the nucleosides. For example, attachment can be at either or both termini of the nucleic acid, or at an internal position, or both. For example, attachment of the label may be done on a ribose of the ribose-phosphate backbone at the 2' or 3' position (the latter for use with terminal labeling), in one embodiment through an amide or amine linkage. Attachment may also be made via a phosphate of the ribose-phosphate backbone, or to the base of a nucleotide. Labels can be attached to one or both ends of a probe or to any one of the nucleotides along the length of a probe.

[0254] Sequencing probes are structured differently depending on the interrogation position desired. For example, in the case of sequencing probes labeled with fluorophores, a single position within each sequencing probe will be correlated with the identity of the fluorophore with which it is labeled. Generally, the fluorophore molecule will be attached to the end of the sequencing probe that is opposite to the end targeted for ligation to the anchor probe.

[0255] By "anchor probe" as used herein is meant an oligonucleotide designed to be complementary to at least a portion of an adaptor, referred to herein as "an anchor site". Adaptors can contain multiple anchor sites for hybridization with multiple anchor probes, as described herein. As discussed further herein, anchor probes of use in the present invention can be designed to hybridize to an adaptor such that at least one end of the anchor probe is flush with one terminus of the adaptor (either "upstream" or "downstream", or both). In further embodiments, anchor probes can be designed to hybridize to at least a portion of an adaptor (a first adaptor site) and also at least one nucleotide of the target nucleic acid adjacent to the adaptor ("overhangs"). As illustrated in Figure 24, anchor probe 2402 comprises a sequence complementary to a portion of the adaptor. Anchor probe 2402 also comprises four degenerate bases at one terminus. This degeneracy allows for a portion of the anchor probe population to fully or partially match the sequence of the target nucleic acid adjacent to the adaptor and allows the anchor probe to hybridize to the adaptor and reach into the target nucleic acid adjacent to the adaptor regardless of the identity of the nucleotides of the target nucleic acid adjacent to the adaptor. This shift of the terminal base of the anchor probe into the target nucleic acid shifts the position of the base to be called closer to the ligation point, thus allowing the fidelity of the ligase to be maintained. In general, ligases ligate probes with higher efficiency if the probes are perfectly complementary to the regions of the target nucleic acid to which they are hybridized, but the fidelity of ligases decreases with distance away from the ligation point. Thus, in order to minimize and/or prevent errors due to incorrect pairing between a sequencing probe and the target nucleic acid, it can be useful to maintain the distance between the nucleotide to be detected and the ligation point of the sequencing and anchor probes. By designing the anchor probe to reach into the target nucleic acid, the fidelity of the ligase is maintained while still allowing a greater number of nucleotides adjacent to each adaptor to be identified. Although the

embodiment illustrated in Figure 24 is one in which the sequencing probe hybridizes to a region of the target nucleic acid on one side of the adaptor, it will be appreciated that embodiments in which the sequencing probe hybridizes on the other side of the adaptor are also encompassed by the invention. In Figure 24, "N" represents a degenerate base and "B" represents nucleotides of undetermined sequence. As will be appreciated, in some embodiments, rather than degenerate bases, universal bases may be used.

[0256] Anchor probes of the invention may comprise any sequence that allows the anchor probe to hybridize to a DNB, generally to an adaptor of a DNB. Such anchor probes may comprise a sequence such that when the anchor probe is hybridized to an adaptor, the entire length of the anchor probe is contained within the adaptor. In some embodiments, anchor probes may comprise a sequence that is complementary to at least a portion of an adaptor and also comprise degenerate bases that are able to hybridize to target nucleic acid regions adjacent to the adaptor. In some exemplary embodiments, anchor probes are hexamers that comprise 3 bases that are complementary to an adaptor and 3 degenerate bases. In some exemplary embodiments, anchor probes are 8-mers that comprise 3 bases that are complementary to an adaptor and 5 degenerate bases. In further exemplary embodiments, particularly when multiple anchor probes are used, a first anchor probe comprises a number of bases complementary to an adaptor at one end and degenerate bases at another end, whereas a second anchor probe comprises all degenerate bases and is designed to ligate to the end of the first anchor probe that comprises degenerate bases. It will be appreciated that these are exemplary embodiments, and that a wide range of combinations of known and degenerate bases can be used to produce anchor probes of use in accordance with the present invention.

[0257] The present invention provides sequencing by ligation methods for identifying sequences of DNBs. In certain aspects, the sequencing by ligation methods of the invention include providing different combinations of anchor probes and sequencing probes, which, when hybridized to adjacent regions on a DNB, can be ligated to form probe ligation products. The probe ligation products are then detected, which provides the identity of one or more nucleotides in the target nucleic acid. By "ligation" as used herein is meant any method of joining two or more nucleotides to each other. Ligation can include chemical as well as enzymatic ligation. In general, the sequencing by ligation methods discussed herein utilize enzymatic ligation by ligases. Such ligases invention can be the same or different than ligases discussed above for creation of the nucleic acid templates. Such ligases include without limitation DNA ligase I, DNA ligase II, DNA ligase III, DNA ligase IV, *E. coli* DNA ligase, T4 DNA ligase, T4 RNA ligase 1, T4 RNA ligase 2, T7 ligase, T3 DNA ligase, and thermostable ligases (including without limitation Taq ligase) and the like. As discussed above, sequencing by ligation methods often rely on the fidelity of ligases to only join probes that are perfectly complementary to the nucleic acid to which they are hybridized. This fidelity will decrease with increasing distance between a base at a particular position in a probe and the ligation point between the two probes. As such, conventional sequencing by ligation methods can be limited in the number of bases that can be identified. The present invention increases the number of bases that can be identified by using multiple probe pools, as is described further herein.

[0258] A variety of hybridization conditions may be used in the sequencing by ligation methods of sequencing as well as other methods of sequencing described herein. These conditions include high, moderate and low stringency conditions; see for example Maniatis et al., Molecular Cloning: A Laboratory Manual, 2d Edition, 1989, and Short Protocols in Molecular Biology, ed. Ausubel, et al, which are hereby

incorporated by reference. Stringent conditions are sequence-dependent and will be different in different circumstances. Longer sequences hybridize specifically at higher temperatures. An extensive guide to the hybridization of nucleic acids is found in Tijssen, Techniques in Biochemistry and Molecular Biology--Hybridization with Nucleic Acid Probes, "Overview of principles of hybridization and the strategy of nucleic acid assays," (1993). Generally, stringent conditions are selected to be about 5-10°C lower than the thermal melting point (Tm) for the specific sequence at a defined ionic strength and pH. The Tm is the temperature (under defined ionic strength, pH and nucleic acid concentration) at which 50% of the probes complementary to the target hybridize to the target sequence at equilibrium (as the target sequences are present in excess, at Tm, 50% of the probes are occupied at equilibrium). Stringent conditions can be those in which the salt concentration is less than about 1.0 M sodium ion, typically about 0.01 to 1.0 M sodium ion concentration (or other salts) at pH 7.0 to 8.3 and the temperature is at least about 30°C for short probes (e.g. 10 to 50 nucleotides) and at least about 60°C for long probes (e.g. greater than 50 nucleotides). Stringent conditions may also be achieved with the addition of helix destabilizing agents such as formamide. The hybridization conditions may also vary when a non-ionic backbone, i.e. PNA is used, as is known in the art. In addition, cross-linking agents may be added after target binding to cross-link, i.e. covalently attach, the two strands of the hybridization complex.

[0259] Although much of the description of sequencing methods is provided in terms of nucleic acid templates of the invention, it will be appreciated that these sequencing methods also encompass identifying sequences in DNBs generated from such nucleic acid templates, as described herein.

[0260] For any of sequencing methods known in the art and described herein using nucleic acid templates of the invention, the present invention provides methods for determining at least about 10 to about 200 bases in target nucleic acids. In further embodiments, the present invention provides methods for determining at least about 20 to about 180, about 30 to about 160, about 40 to about 140, about 50 to about 120, about 60 to about 100, and about 70 to about 80 bases in target nucleic acids. In still further embodiments, sequencing methods are used to identify at least 5, 10, 15, 20, 25, 30 or more bases adjacent to one or both ends of each adaptor in a nucleic acid template of the invention.

[0261] Any of the sequencing methods described herein and known in the art can be applied to nucleic acid templates and/or DNBs of the invention in solution or to nucleic acid templates and/or DNBs disposed on a surface and/or in an array.

VIIB(i). Single cPAL

[0262] In one aspect, the present invention provides methods for identifying sequences of DNBs by using combinations of sequencing and anchor probes that hybridize to adjacent regions of a DNB and are ligated, usually by application of a ligase. Such methods are generally referred to herein as cPAL (combinatorial probe anchor ligation) methods. In one aspect, cPAL methods of the invention produce probe ligation products comprising a single anchor probe and a single sequencing probe. Such cPAL methods in which only a single anchor probe is used are referred to herein as “single cPAL”.

[0263] One embodiment of single cPAL is illustrated in Figure 23. A monomeric unit 2301 of a DNB comprises a target nucleic acid 2309 and an adaptor 2308. An anchor probe 2302 hybridizes to a complementary region on adaptor 2308. In the example illustrated in Figure 23, anchor probe 2302 hybridizes to the adaptor region directly adjacent to target nucleic acid 2309, although, as is discussed

further herein, anchor probes can also be designed to reach into the target nucleic acid adjacent to an adaptor by incorporating a desired number of degenerate bases at the terminus of the anchor probe. A pool of differentially labeled sequencing probes 2306 will hybridize to complementary regions of the target nucleic acid. A sequencing probe 2310 that hybridizes to the region of target nucleic acid 2309 adjacent to anchor probe 2302 will be ligated to the anchor probe form a probe ligation product. The efficiency of hybridization and ligation is increased when the base in the interrogation position of the probe is complementary to the unknown base in the detection position of the target nucleic acid. This increased efficiency favors ligation of perfectly complementary sequencing probes to anchor probes over mismatch sequencing probes. As discussed above, ligation is generally accomplished enzymatically using a ligase, but other ligation methods can also be utilized in accordance with the invention. In Figure 23, "N" represents a degenerate base and "B" represents nucleotides of undetermined sequence. As will be appreciated, in some embodiments, rather than degenerate bases, universal bases may be used.

[0264] As also discussed above, the sequencing probes can be oligonucleotides representing each base type at a specific position and labeled with a detectable label that differentiates each sequencing probe from the sequencing probes with other nucleotides at that position. Thus, in the example illustrated in Figure 23, a sequencing probe 2310 that hybridizes adjacent to anchor probe 2302 and is ligated to the anchor probe will identify the base at a position in the target nucleic acid 5 bases from the adaptor as a "G". Multiple cycles of anchor probe and sequencing probe hybridization and ligation can be used to identify a desired number of bases of the target nucleic acid on each side of each adaptor in a DNB.

[0265] As will be appreciated, hybridization of the anchor probe and the sequencing probe can be sequential or simultaneous in any of the cPAL methods described herein.

[0266] In the embodiment illustrated in Figure 23, sequencing probe 2310 hybridizes to a region "upstream" of the adaptor, however it will be appreciated that sequencing probes may also hybridize "downstream" of the adaptor. The terms "upstream" and "downstream" refer to the regions 5' and 3' of the adaptor, depending on the orientation of the system. In general, "upstream" and "downstream" are relative terms and are not meant to be limiting; rather they are used for ease of understanding. As illustrated in Figure 6, a sequencing probe 607 can hybridize downstream of adaptor 604 to identify a nucleotide 4 bases away from the interface between the adaptor and the target nucleic acid 603. In further embodiments, sequencing probes can hybridize both upstream and downstream of the adaptor to identify nucleotides at positions in the nucleic acid on both sides of the adaptor. Such embodiments allow generation of multiple points of data from each adaptor for each hybridization-ligation-detection cycle of the single cPAL method.

[0267] In some embodiments, probes used in a single cPAL method may have from about 3 to about 20 bases corresponding to an adaptor and from about 1 to about 20 degenerate bases (i.e., in a pool of anchor probes). Such anchor probes may also include universal bases, as well as combinations of degenerate and universal bases.

[0268] In some embodiments, anchor probes with degenerated bases may have about 1-5 mismatches with respect to the adaptor sequence to increase the stability of full match hybridization at the degenerated bases. Such a design provides an additional way to control the stability of the ligated anchor and sequencing probes to favor those probes that are perfectly matched to the target (unknown) sequence. In further embodiments, a number of bases in the degenerate portion of the anchor probes

may be replaced with abasic sites (i.e., sites which do not have a base on the sugar) or other nucleotide analogs to influence the stability of the hybridized probe to favor the full match hybrid at the distal end of the degenerate part of the anchor probe that will participate in the ligation reactions with the sequencing probes, as described herein. Such modifications may be incorporated, for example, at interior bases, particularly for anchor probes that comprise a large number (i.e., greater than 5) of degenerated bases. In addition, some of the degenerated or universal bases at the distal end of the anchor probe may be designed to be cleavable after hybridization (for example by incorporation of a uracil) to generate a ligation site to the sequencing probe or to a second anchor probe, as described further below.

[0269] In further embodiments, the hybridization of the anchor probes can be controlled through manipulation of the reaction conditions, for example the stringency of hybridization. In an exemplary embodiment, the anchor hybridization process may start with conditions of high stringency (higher temperature, lower salt, higher pH, higher concentration of formamide, and the like), and these conditions may be gradually or stepwise relaxed. This may require consecutive hybridization cycles in which different pools of anchor probes are removed and then added in subsequent cycles. Such methods provide a higher percentage of target nucleic acid occupied with perfectly complementary anchor probes, particularly anchor probes perfectly complementary at positions at the distal end that will be ligated to the sequencing probe. Hybridization time at each stringency condition may also be controlled to obtain greater numbers of full match hybrids.

VIIB(ii). Double cPAL (and beyond)

[0270] In still further embodiments, the present invention provides cPAL methods utilizing two ligated anchor probes in every hybridization-ligation cycle. See for example U.S. Patent Application Serial Nos. 60/992,485; 61/026,337; 61/035,914 and 61/061,134**,** which are hereby expressly incorporated by reference in their entirety, and especially the examples and claims. Figure 25 illustrates an example of a "double cPAL" method in which a first anchor probe 2502 and a second anchor probe 2505 hybridize to complimentary regions of an adaptor; that is, the first anchor probe hybridizes to the first anchor site and the second anchor probe hybridizes to the second adaptor site. In the example illustrated in Figure 25, the first anchor probe 2502 is fully complementary to a region of the adaptor 2511 (the first anchor site), and the second anchor probe 2505 is complementary to the adaptor region adjacent to the hybridization position of the first anchor probe (the second anchor site). In general, the first and second anchor sites are adjacent.

[0271] The second anchor probe may optionally also comprises degenerate bases at the terminus that is not adjacent to the first anchor probe such that it will hybridize to a region of the target nucleic acid 2512 adjacent to adaptor 2511. This allows sequence information to be generated for target nucleic acid bases farther away from the adaptor/target interface. Again, as outlined herein, when a probe is said to have "degenerate bases", it means that the probe actually comprises a set of probes, with all possible combinations of sequences at the degenerate positions. For example, if an anchor probe is 9 bases long with 6 known bases and three degenerate bases, the anchor probe is actually a pool of 64 probes.

[0272] The second anchor probe is generally too short to be maintained alone in its duplex hybridization state, but upon ligation to the first anchor probe it forms a longer anchor probe that is stable for subsequent methods. In the some embodiments, the second anchor probe has about 1 to about 5 bases

that are complementary to the adaptor and about 5 to about 10 bases of degenerate sequence. As discussed above for the "single cPAL" method, a pool of sequencing probes 2508 representing each base type at a detection position of the target nucleic acid and labeled with a detectable label that differentiates each sequencing probe from the sequencing probes with other nucleotides at that position is hybridized 2509 to the adaptor-anchor probe duplex and ligated to the terminal 5' or 3' base of the ligated anchor probes. In the example illustrated in Figure 25, the sequencing probes are designed to interrogate the base that is five positions 5' of the ligation point between the sequencing probe 2514 and the ligated anchor probes 2513. Since the second anchor probe 2505 has five degenerate bases at its 5' end, it reaches 5 bases into the target nucleic acid 2512, allowing interrogation with the sequencing probe at a full 10 bases from the interface between the target nucleic acid 2512 and the adaptor 2511. In Figure 25, "N" represents a degenerate base and "B" represents nucleotides of undetermined sequence. As will be appreciated, in some embodiments, rather than degenerate bases, universal bases may be used.

[0273] In some embodiments, the second anchor probe may have about 5-10 bases corresponding to an adaptor and about 5-15 bases, which are generally degenerated, corresponding to the target nucleic acid. This second anchor probe may be hybridized first under optimal conditions to favor high percentages of target occupied with full match at a few bases around the ligation point between the two anchor probes. The first adaptor probe and/or the sequencing probe may be hybridized and ligated to the second anchor probe in a single step or sequentially. In some embodiments, the first and second anchor probes may have at their ligation point from about 5 to about 50 complementary bases that are not complementary to the adaptor, thus forming a "branching-out" hybrid. This design allows an adaptor-specific stabilization of the hybridized second anchor probe. In some embodiments, the second anchor probe is ligated to the sequencing probe before hybridization of the first anchor probe; in some embodiments the second anchor probe is ligated to the first anchor probe prior to hybridization of the sequencing probe; in some embodiments the first and second anchor probes and the sequencing probe hybridize simultaneously and ligation occurs between the first and second anchor probe and between the second anchor probe and the sequencing probe simultaneously or essentially simultaneously, while in other embodiments the ligation between the first and second anchor probe and between the second anchor probe and the sequencing probe occurs sequentially in any order. Stringent washing conditions can be used to remove unligated probes; (e.g., using temperature, pH, salt, a buffer with an optimal concentration of formamide can all be used, with optimal conditions and/or concentrations being determined using methods known in the art). Such methods can be particularly useful in methods utilizing second anchor probes with large numbers of degenerated bases that are hybridized outside of the corresponding junction point between the anchor probe and the target nucleic acid.

[0274] In certain embodiments, double cPAL methods utilize ligation of two anchor probes in which one anchor probe is fully complementary to an adaptor and the second anchor probe is fully degenerate (again, actually a pool of probes). An example of such a double cPAL method is illustrated in Figure 26, in which the first anchor probe 2602 is hybridized to adaptor 2611 of DNB 2601. The second anchor probe 2605 is fully degenerate and is thus able to hybridize to the unknown nucleotides of the region of the target nucleic acid 2612 adjacent to adaptor 2611. The second anchor probe is designed to be too short to be maintained alone in its duplex hybridization state, but upon ligation to the first anchor probe the formation of the longer ligated anchor probe construct provides the stability needed for subsequent

steps of the cPAL process. The second fully degenerate anchor probe may in some embodiments be from about 5 to about 20 bases in length. For longer lengths (i.e., above 10 bases), alterations to hybridization and ligation conditions may be introduced to lower the effective Tm of the degenerate anchor probe. The shorter second anchor probe will generally bind non-specifically to target nucleic acid and adaptors, but its shorter length will affect hybridization kinetics such that in general only those second anchor probes that are perfectly complementary to regions adjacent to the adaptors and the first anchor probes will have the stability to allow the ligase to join the first and second anchor probes, generating the longer ligated anchor probe construct. Non-specifically hybridized second anchor probes will not have the stability to remain hybridized to the DNB long enough to subsequently be ligated to any adjacently hybridized sequencing probes. In some embodiments, after ligation of the second and first anchor probes, any unligated anchor probes will be removed, usually by a wash step. In Figure 26, "N" represents a degenerate base and "B" represents nucleotides of undetermined sequence. As will be appreciated, in some embodiments, rather than degenerate bases, universal bases may be used.

[0275] In further exemplary embodiments, the first anchor probe will be a hexamer comprising 3 bases complementary to the adaptor and 3 degenerate bases, whereas the second anchor probe comprises only degenerate bases and the first and second anchor probes are designed such that only the end of the first anchor probe with the degenerate bases will ligate to the second anchor probe. In further exemplary embodiments, the first anchor probe is an 8-mer comprising 3 bases complementary to an adaptor and 5 degenerate bases, and again the first and second anchor probes are designed such that only the end of the first anchor probe with the degenerate bases will ligate to the second anchor probe. It will be appreciated that these are exemplary embodiments and that a wide range of combinations of known and degenerate bases can be used in the design of both the first and second (and in some embodiments the third and/or fourth) anchor probes.

[0276] In variations of the above described examples of a double cPAL method, if the first anchor probe terminates closer to the end of the adaptor, the second anchor probe will be proportionately more degenerate and therefore will have a greater potential to not only ligate to the end of the first anchor probe but also to ligate to other second anchor probes at multiple sites on the DNB. To prevent such ligation artifacts, the second anchor probes can be selectively activated to engage in ligation to a first anchor probe or to a sequencing probe. Such activation include selectively modifying the termini of the anchor probes such that they are able to ligate only to a particular anchor probe or sequencing probe in a particular orientation with respect to the adaptor. For example, 5' and 3' phosphate groups can be introduced to the second anchor probe, with the result that the modified second anchor probe would be able to ligate to the 3' end of a first anchor probe hybridized to an adaptor, but two second anchor probes would not be able to ligate to each other (because the 3' ends are phosphorylated, which would prevent enzymatic ligation). Once the first and second anchor probes are ligated, the 3' ends of the second anchor probe can be activated by removing the 3' phosphate group (for example with T4 polynucleotide kinase or phosphatases such as shrimp alkaline phosphatase and calf intestinal phosphatase).

[0277] If it is desired that ligation occur between the 3' end of the second anchor probe and the 5' end of the first anchor probe, the first anchor probe can be designed and/or modified to be phosphorylated on its 5' end and the second anchor probe can be designed and/or modified to have no 5' or 3' phosphorylation. Again, the second anchor probe would be able to ligate to the first anchor probe, but not to other second anchor probes. Following ligation of the first and second anchor probes, a 5'

phosphate group can be produced on the free terminus of the second anchor probe (for example, by using T4 polynucleotide kinase) to make it available for ligation to sequencing probes in subsequent steps of the cPAL process.

[0278] In some embodiments, the two anchor probes are applied to the DNBs simultaneously. In some embodiments, the two anchor probes are applied to the DNBs sequentially, allowing one of the anchor probes to hybridize to the DNBs before the other. In some embodiments, the two anchor probes are ligated to each other before the second adaptor is ligated to the sequencing probe. In some embodiments, the anchor probes and the sequencing probe are ligated in a single step. In embodiments in which two anchor probes and the sequencing probe are ligated in a single step, the second adaptor can be designed to have enough stability to maintain its position until all three probes (the two anchor probes and the sequencing probe) are in place for ligation. For example, a second anchor probe comprising five bases complementary to the adaptor and five degenerate bases for hybridization to the region of the target nucleic acid adjacent to the adaptor can be used. Such a second anchor probe may have sufficient stability to be maintained with low stringency washing, and thus a ligation step would not be necessary between the steps of hybridization of the second anchor probe and hybridization of a sequencing probe. In the subsequent ligation of the sequencing probe to the second anchor probe, the second anchor probe would also be ligated to the first anchor probe, resulting in a duplex with increased stability over any of the anchor probes or sequencing probes alone.

[0279] Similar to the double cPAL method described above, it will be appreciated that cPAL with three or more anchor probes is also encompassed by the present invention. Such anchor probes can be designed in accordance with methods described herein and known in the art to hybridize to regions of adaptors such that one terminus of one of the anchor probes is available for ligation to sequencing probes hybridized adjacent to the terminal anchor probe. In an exemplary embodiment, three anchor probes are provided – two are complementary to different sequences within an adaptor and the third comprises degenerate bases to hybridize to sequences within the target nucleic acid. In a further embodiment, one of the two anchors complementary to sequences within the adaptor may also comprise one or more degenerate bases at on terminus, allowing that anchor probe to reach into the target nucleic acid for ligation with the third anchor probe. In further embodiments, one of the anchor probes may be fully or partially complementary to the adaptor and the second and third anchor probes will be fully degenerate for hybridization to the target nucleic acid. Four or more fully degenerate anchor probes can in further embodiments be ligated sequentially to the three ligated anchor probes to achieve extension of reads further into the target nucleic acid sequence. In an exemplary embodiment, a first anchor probe comprising twelve bases complementary to an adaptor may ligate with a second hexameric anchor probe in which all six bases are degenerate. A third anchor, also a fully degenerate hexamer, can also ligate to the second anchor probe to further extend into the unknown sequence of the target nucleic acid. A fourth, fifth, sixth, etc. anchor probe may also be added to extend even further into the unknown sequence. In still further embodiments and in accordance with any of the cPAL methods described herein, one or more of the anchor probes may comprise one or more labels that serve to “tag” the anchor probe and/or identify the particular anchor probe hybridized to an adaptor of a DNB.

VIIB(iii). Detecting fluorescently labeled sequencing probes

[0280] As discussed above, sequencing probes used in accordance with the present invention may be detectably labeled with a wide variety of labels. Although the following description is primarily directed to embodiments in which the sequencing probes are labeled with fluorophores, it will be appreciated that similar embodiments utilizing sequencing probes comprising other kinds of labels are encompassed by the present invention.

[0281] Multiple cycles of cPAL (whether single, double, triple, etc.) will identify multiple bases in the regions of the target nucleic acid adjacent to the adaptors. In brief, the cPAL methods are repeated for interrogation of multiple bases within a target nucleic acid by cycling anchor probe hybridization and enzymatic ligation reactions with sequencing probe pools designed to detect nucleotides at varying positions removed from the interface between the adaptor and target nucleic acid. In any given cycle, the sequencing probes used are designed such that the identity of one or more of bases at one or more positions is correlated with the identity of the label attached to that sequencing probe. Once the ligated sequencing probe (and hence the base(s) at the interrogation position(s) is detected, the ligated complex is stripped off of the DNB and a new cycle of adaptor and sequencing probe hybridization and ligation is conducted.

[0282] In general, four fluorophores are generally used to identify a base at an interrogation position within a sequencing probe, and a single base is queried per hybridization-ligation-detection cycle. However, as will be appreciated, embodiments utilizing 8, 16, 20 and 24 fluorophores or more are also encompassed by the present invention. Increasing the number of fluorophores increases the number of bases that can be identified during any one cycle.

[0283] In one exemplary embodiment, a set of 7-mer pools of sequencing probes is employed having the following structures:

3'-F1-NNNNNNAp
3'-F2-NNNNNNGp
3'-F3-NNNNNNCp
3'-F4-NNNNNNTp

[0284] The "p" represents a phosphate available for ligation and "N" represents degenerate bases. F1-F4 represent four different fluorophores – each fluorophore is thus associated with a particular base. This exemplary set of probes would allow detection of the base immediately adjacent to the adaptor upon ligation of the sequencing probe to an anchor probe hybridized to the adaptor. To the extent that the ligase used to ligate the sequencing probe to the anchor probe discriminates for complementarity between the base at the interrogation position of the probe and the base at the detection position of the target nucleic acid, the fluorescent signal that would be detected upon hybridization and ligation of the sequencing probe provides the identity of the base at the detection position of the target nucleic acid.

[0285] In some embodiments, a set of sequencing probes will comprise three differentially labeled sequencing probes, with a fourth optional sequencing probe left unlabeled.

[0286] After performing a hybridization-ligation-detection cycle, the anchor probe-sequencing probe ligation products are stripped and a new cycle is begun. In some embodiments, accurate sequence information can be obtained as far as six bases or more from the ligation point between the anchor and

sequencing probes and as far as twelve bases or more from the interface between the target nucleic acid and the adaptor. The number of bases that can be identified can be increased using methods described herein, including the use of anchor probes with degenerate ends that are able to reach further into the target nucleic acid.

[0287] Imaging acquisition may be performed using methods known in the art, including the use of commercial imaging packages such as Metamorph (Molecular Devices, Sunnyvale, CA). 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.

[0288] In an exemplary embodiment, DNBs disposed on a surface undergo a cycle of cPAL as described herein in which the sequencing probes utilized are labeled with four different fluorophores (each corresponding to a particular base at an interrogation position within the probe). To determine the identity of a base of each DNB disposed on the surface, each field of view ("frame") is imaged with four different wavelengths corresponding the to the four fluorescently labeled sequencing probes. All images from each cycle are saved in a cycle directory, where the number of images is four times the number of frames (when four fluorophores are used). Cycle image data can then be saved into a directory structure organized for downstream processing.

[0289] In some embodiments, data extraction will rely on two types of image data: bright-field images to demarcate the positions of all DNBs on a surface, and sets of fluorescence images acquired during each sequencing cycle. Data extraction software can be used to identify all objects with the bright-field images and then for each such object, the software can be used to compute 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 data points (also referred to herein as "base calls") are consolidated, yielding a discontinuous sequencing read for each DNB.

[0290] The population of identified bases can then be assembled to provide sequence information for the target nucleic acid and/or identify the presence of particular sequences in the target nucleic acid. In some embodiments, the identified bases are assembled into a complete sequence through alignment of overlapping sequences obtained from multiple sequencing cycles performed on multiple DNBs. As used herein, the term "complete sequence" refers to the sequence of partial or whole genomes as well as partial or whole target nucleic acids. In further embodiments, assembly methods utilize algorithms that can be used to "piece together" overlapping sequences to provide a complete sequence. In still further embodiments, reference tables are used to assist in assembling the identified sequences into a complete sequence. 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.

[0291] In any of the embodiments of the invention discussed herein, a population of nucleic acid templates and/or DNBs may comprise a number of target nucleic acids to substantially cover a whole genome or a whole target polynucleotide. As used herein, "substantially covers" means that the amount of nucleotides (i.e., target sequences) analyzed contains an equivalent of at least two copies of the target polynucleotide, or in another aspect, at least ten copies, or in another aspect, at least twenty copies, or in another aspect, at least 100 copies. Target polynucleotides may include DNA fragments, including genomic DNA fragments and cDNA fragments, and RNA fragments. Guidance for the step of reconstructing target polynucleotide sequences can be found in the following references, which are incorporated by reference: Lander et al, Genomics, 2: 231-239 (1988); Vingron et al, J. Mol. Biol., 235: 1-12 (1994); and like references.

VIIB(iv). Sets of probes

[0292] As will be appreciated, different combinations of sequencing and anchor probes can be used in accordance with the various cPAL methods described above. The following descriptions of sets of probes (also referred to herein as "pools of probes") of use in the present invention are exemplary embodiments and it will be appreciated that the present invention is not limited to these combinations.

[0293] In one aspect, sets of probes are designed for identification of nucleotides at positions at a specific distance from an adaptor. For example, certain sets of probes can be used to identify bases up to 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30 and more positions away from the adaptor. As discussed above, anchor probes with degenerate bases at one terminus can be designed to reach into the target nucleic acid adjacent to an adaptor, allowing sequencing probes to ligate further away from the adaptor and thus provide the identity of a base further away from the adaptor.

[0294] In an exemplary embodiment, a set of probes comprises at least two anchor probes designed to hybridize to adjacent regions of an adaptor. In one embodiment, the first anchor probe is fully complementary to a region of the adaptor, while the second anchor probe is complementary to the adjacent region of the adaptor. In some embodiments, the second anchor probe will comprise one or more degenerate nucleotides that extend into and hybridize to nucleotides of the target nucleic acid adjacent to the adaptor. In an exemplary embodiment, the second anchor probe comprises at least 1-10 degenerate bases. In a further exemplary embodiment, the second anchor probe comprises 2-9, 3-8, 4-7, and 5-6 degenerate bases. In a still further exemplary embodiment, the second anchor probe comprises one or more degenerate bases at one or both termini and/or within an interior region of its sequence.

[0295] In a further embodiment, a set of probes will also comprise one or more groups of sequencing probes for base determination in one or more detection positions with a target nucleic acid. In one embodiment, the set comprises enough different groups of sequencing probes to identify about 1 to about 20 positions within a target nucleic acid. In a further exemplary embodiment, the set comprises enough groups of sequencing probes to identify about 2 to about 18, about 3 to about 16, about 4 to about 14, about 5 to about 12, about 6 to about 10, and about 7 to about 8 positions within a target nucleic acid.

[0296] In further exemplary embodiments, 10 pools of labeled or tagged probes will be used in accordance with the invention. In still further embodiments, sets of probes will include two or more anchor probes with different sequences. In yet further embodiments, sets of probes will include 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15 or more anchor probes with different sequences.

[0297] In a further exemplary embodiment, a set of probes is provided comprising one or more groups of sequencing probes and three anchor probes. The first anchor probe is complementary to a first region of an adaptor, the second anchor probe is complementary to a second region of an adaptor, and the second region and the first region are adjacent to each other. The third anchor probe comprises three or more degenerate nucleotides and is able to hybridize to nucleotides in the target nucleic acid adjacent to the adaptor. The third anchor probe may also in some embodiments be complementary to a third region of the adaptor, and that third region may be adjacent to the second region, such that the second anchor probe is flanked by the first and third anchor probes.

[0298] In some embodiments, sets of anchor and/or sequencing probes will comprise variable concentrations of each type of probe, and the variable concentrations may in part depend on the degenerate bases that may be contained in the anchor probes. For example, probes that will have lower hybridization stability, such as probes with greater numbers of A's and/or T's, can be present in higher relative concentrations as a way to offset their lower stabilities. In further embodiments, these differences in relative concentrations are established by preparing smaller pools of probes independently and then mixing those independently generated pools of probes in the proper amounts.

VIIB(v). Other sequencing methods

[0299] In one aspect, methods and compositions of the present invention are used in combination with techniques such as those described in WO2007120208, WO2006073504, WO2007133831, and US2007099208, and U.S. Patent Application Nos. 60/992,485; 61/026,337; 61/035,914; 61/061,134; 61/116,193; 61/102,586; 12/265,593; 12/266,385; 11/938,096; 11/981,804; No 11/981,797; 11/981,793; 11/981,767; 11/981,761; 11/981,730; 11/981,685; 11/981,661; 11/981,607; 11/981,605; 11/927,388; 11/927,356; 11/679,124; 11/541,225; 10/547,214; 11/451,692; and 11/451,691, all of which are incorporated herein by reference in their entirety for all purposes and in particular for all teachings related to sequencing, particularly sequencing of concatemers.

[0300] In a further aspect, sequences of DNBs are identified using sequencing methods known in the art, including, but not limited to, hybridization-based methods, such as disclosed in Drmanac, U.S. patents 6,864,052; 6,309,824; and 6,401,267; and Drmanac et al, U.S. patent publication 2005/0191656, and sequencing by synthesis methods, e.g. Nyren et al, U.S. patent 6,210,891; Ronaghi, U.S. patent 6,828,100; Ronaghi et al (1998), Science, 281: 363-365; Balasubramanian, U.S. patent 6,833,246; Quake, U.S. patent 6,911,345; Li et al, Proc. Natl. Acad. Sci., 100: 414-419 (2003); Smith et al, PCT publication WO 2006/074351; and ligation-based methods, e.g. Shendure et al (2005), Science, 309: 1728-1739, Macevicz, U.S. patent 6,306,597, wherein each of these references is herein incorporated by reference in its entirety for all purposes and in particular teachings regarding the figures, legends and accompanying text describing the compositions, methods of using the compositions and methods of making the compositions, particularly with respect to sequencing.

[0301] In some embodiments, nucleic acid templates of the invention, as well as DNBs generated from those templates, are used in sequencing by synthesis methods. The efficiency of sequencing by synthesis methods utilizing nucleic acid templates of the invention is increased over conventional sequencing by synthesis methods utilizing nucleic acids that do not comprise multiple interspersed adaptors. Rather than a single long read, nucleic acid templates of the invention allow for multiple short reads that each start at one of the adaptors in the template. Such short reads consume fewer labeled dNTPs, thus saving on the cost of reagents. In addition, sequencing by synthesis reactions can be performed on DNB arrays, which provide a high density of sequencing targets as well as multiple copies of monomeric units. Such arrays provide detectable signals at the single molecule level while at the same time providing an increased amount of sequence information, because most or all of the DNB monomeric units will be extended without losing sequencing phase. The high density of the arrays also reduces reagent costs – in some embodiments the reduction in reagent costs can be from about 30 to about 40% over conventional sequencing by synthesis methods. In some embodiments, the interspersed adaptors of the nucleic acid templates of the invention provide a way to combine about two to about ten standard reads if inserted at distances of from about 30 to about 100 bases apart from one another. In such embodiments, the newly synthesized strands will not need to be stripped off for further sequencing cycles, thus allowing the use of a single DNB array through about 100 to about 400 sequencing by synthesis cycles.

VIIC. Two-phase sequencing

[0302] In one aspect, the present invention provides methods for "two-phase" sequencing, which is also referred to herein as "shotgun sequencing". Such methods are described in U.S. Patent Application No. 12/325,922, filed December 1, 2008, which is hereby incorporated by reference in its entirety for all purposes and in particular for all teachings related to two-phase or shotgun sequencing.

[0303] Generally, two phase-sequencing methods of use in the present invention comprise the following steps: (a) sequencing the target nucleic acid to produce a primary target nucleic acid sequence that comprises one or more sequences of interest; (b) synthesizing a plurality of target-specific oligonucleotides, wherein each of said plurality of target-specific oligonucleotides corresponds to at least one of the sequences of interest; (c) providing a library of fragments of the target nucleic acid (or constructs that comprise such fragments and that may further comprise, for example, adaptors and other sequences as described herein) that hybridize to the plurality of target-specific oligonucleotides; and (d) sequencing the library of fragments (or constructs that comprise such fragments) to produce a secondary target nucleic acid sequence. In order to close gaps due to missing sequence or resolve low confidence base calls in a primary sequence of genomic DNA, such as human genomic DNA, the number of target-specific oligonucleotides that are synthesized for these methods may be from about ten thousand to about one million; thus the present invention contemplates the use of at least about 10,000 target-specific oligonucleotides, or about 25,000, or about 50,000, or about 100,000, or about 20,000, or about 50,000, or about 100,000, or about 200,000 or more.

[0304] In saying that the plurality of target-specific oligonucleotides "corresponds to" at least one of the sequences of interest, it is meant that such target-specific oligonucleotides are designed to hybridize to the target nucleic acid in proximity to, including but not limited to, adjacent to, the sequence of interest

such that there is a high likelihood that a fragment of the target nucleic acid that hybridizes to such an oligonucleotides will include the sequence of interest. Such target-specific oligonucleotides are therefore useful for hybrid capture methods to produce a library of fragments enriched for such sequences of interest, as sequencing primers for sequencing the sequence of interest, as amplification primers for amplifying the sequence of interest, or for other purposes.

[0305] In shotgun sequencing and other sequencing methods according to the present invention, after assembly of sequencing reads, to the skilled person it is apparent from the assembled sequence that gaps exist or that there is low confidence in one or more bases or stretches of bases at a particular site in the sequence. Sequences of interest, which may include such gaps, low confidence sequence, or simply different sequences at a particular location (i.e., a change of one or more nucleotides in target sequence), can also be identified by comparing the primary target nucleic acid sequence to a reference sequence.

[0306] According to one embodiment of such methods sequencing the target nucleic acid to produce a primary target nucleic acid sequence comprises computerized input of sequence readings and computerized assembly of the sequence readings to produce the primary target nucleic acid sequence. In addition, design of the target-specific oligonucleotides can be computerized, and such computerized synthesis of the target-specific oligonucleotides can be integrated with the computerized input and assembly of the sequence readings and design of the target-specific oligonucleotides. This is especially helpful since the number of target-specific oligonucleotides to be synthesized can be in the tens of thousands or hundreds of thousands for genomes of higher organisms such as humans, for example. Thus the invention provides automated integration of the process of creating the oligonucleotide pool from the determined sequences and the regions identified for further processing. In some embodiments, a computer-driven program uses the identified regions and determined sequence near or adjacent to such identified regions to design oligonucleotides to isolate and/or create new fragments that cover these regions. The oligonucleotides can then be used as described herein to isolate fragments, either from the first sequencing library, from a precursor of the first sequencing library, from a different sequencing library created from the same target nucleic acid, directly from target nucleic acids, and the like. In further embodiments, this automated integration of identifying regions for further analysis and isolating/creating the second library defines the sequence of the oligonucleotides within the oligonucleotide pool and directs synthesis of these oligonucleotides.

[0307] In some embodiments of the two phase sequencing methods of the invention, a releasing process is performed after the hybrid capture process, and in other aspects of the technology, an amplification process is performed before the second sequencing process.

[0308] In still further embodiments, some or all regions are identified in the identifying step by comparison of determined sequences with a reference sequence. In some aspects, the second shotgun sequencing library is isolated using a pool of oligonucleotides comprising oligonucleotides based on a reference sequence. Also, in some aspects, the pool of oligonucleotides comprises at least 1000 oligonucleotides of different sequence, in other aspects, the pool of oligonucleotides comprises at least 10,000, 25,000, 50,000, 75,000, or 100,000 or more oligonucleotides of different sequence

[0309] In some aspects of the invention, one or more of the sequencing processes used in this two-phase sequencing method is performed by sequencing-by-ligation, and in other aspects, one or more of the sequencing processes is performed by sequencing-by-hybridization or sequencing-by-synthesis.

[0310] In certain aspects of the invention, between about 1 to about 30% of the complex target nucleic acid is identified as having to be re-sequenced in Phase II of the methods, and in other aspects, between about 1 to about 10% of the complex target nucleic acid is identified as having to be re-sequenced in Phase II of the methods. In some aspects, coverage for the identified percentage of complex target nucleic acid is between about 25x to about 100x.

[0311] In further aspects, 1 to about 10 target-specific selection oligonucleotides are defined and synthesized for each target nucleic acid region that is re-sequenced in Phase II of the methods; in other aspects, about 3 to about 6 target-specific selection oligonucleotides are defined for each target nucleic acid region that is re-sequenced in Phase II of the methods.

[0312] In still further aspects of the technology, the target-specific selection oligonucleotides are identified and synthesized by an automated process, wherein the process that identifies regions of the complex nucleic acid missing nucleic acid sequence or having low confidence nucleic acid sequence and defines sequences for the target-specific selection oligonucleotides communicates with oligonucleotide synthesis software and hardware to synthesize the target-specific selection oligonucleotides. In other aspects of the technology, the target-specific selection oligonucleotides are between about 20 and about 30 bases in length, and in some aspects are unmodified.

[0313] Not all regions identified for further analysis may actually exist in the complex target nucleic acid. One reason for predicted lack of coverage in a region may be that a region expected to be in the complex target nucleic acid may actually not be present (*e.g.,* a region may be deleted or re-arranged in the target nucleic acid), and thus not all oligonucleotides produced from the pool may isolate a fragment for inclusion in the second shotgun sequencing library. In some embodiments, at least one oligonucleotide will be designed and created for each region identified for further analysis. In further embodiments, an average of three or more oligonucleotides will be provided for each region identified for further analysis. It is a feature of the invention that the pool of oligonucleotides can be used directly to create the second shotgun sequencing library by polymerase extension of the oligonucleotides using templates derived from a target nucleic acid. It is another feature of the invention that the pool of oligonucleotides can be used directly to create amplicons via circle dependent replication using the oligonucleotide pools and circle dependent replication. It is another feature of the invention that the methods will provide sequencing information to identify absent regions of interest, *e.g.* predicted regions that were identified for analysis but which do not exist, *e.g.,* due to a deletion or rearrangement.

[0314] The above described embodiments of the two-phase sequencing method can be used in combination with any of the nucleic acid constructs and sequencing methods described herein and known in the art.

VIID. SNP detection

[0315] Methods and compositions discussed above can in further embodiments be used to detect specific sequences in nucleic acid constructs such as DNBs. In particular, cPAL methods utilizing sequencing and anchor probes can be used to detect polymorphisms or sequences associated with a genetic mutation, including single nucleotide polymorphisms (SNPs). For example, to detect the presence of a SNP, two sets of differentially labeled sequencing probes can be used, such that detection of one probe over the other indicates whether a polymorphism present in the sample. Such sequencing

probes can be used in conjunction with anchor probes in methods similar to the cPAL methods described above to further improve the specificity and efficiency of detection of the SNP.

VIII. Exemplary embodiments

[0316] In one aspect, the present invention provides a method for determining a sequence in a target nucleic acid. This method includes the steps of: (a) providing a sequencing template that includes a fragment of the target nucleic acid and an adaptor that includes at least a first anchor site; (b) hybridizing an anchor probe to the anchor site, and the anchor probe includes a region complementary to the adaptor site and three or more degenerate bases for binding in the target nucleic acid sequence; (c) hybridizing a pool of sequencing probes for determination of the sequence of one or more nucleotides in a defined position relative to the adaptor, wherein the sequencing probe is detectably labeled to identify the presence of a particular base; (d) ligating the anchor probe and the sequencing probe; and (e) detecting the sequencing probe, thereby determining a sequence in the target nucleic acid.

[0317] In a further aspect and in accordance with any of the above, the invention provides a method of determining the identification of a first nucleotide at a detection position of a target sequence comprising a plurality of detection positions. This method includes the steps of: (a) providing a surface with a plurality of concatemers, wherein each concatemer includes a plurality of monomers and each monomer comprises: (i) a first target domain of the target sequence comprising a first set of target detection positions; (ii) at least a first adaptor comprising: (1) a first anchor site; and (2) a second adjacent anchor site; (b) hybridizing a first anchor probe to the first anchor site; (c) hybridizing a second anchor probe to the second anchor site, wherein the second anchor probe also hybridizes to sequences outside the second anchor site; (d) hybridizing at least a first sequencing probe to the first target domain, wherein the first sequencing probe comprises: (i) a first probe domain complementary to the target domain; (ii) a unique nucleotide at a first interrogation position; and (iii) a label; under conditions wherein if the unique nucleotide is complementary to the first nucleotide, the sequencing probe hybridizes to the concatemer; (e) ligating the anchor probes and the sequencing probe; and (f) identifying the first nucleotide.

[0318] In one embodiment, and in accordance with any of the above, the present invention provides a method of determining the identification of a first nucleotide at a detection position of a target sequence in which a set of sequencing probes are contacted with a surface comprising a plurality of concatemers. In this embodiment, each sequencing probe includes: (a) a first probe domain complementary to the target domain; (b) a unique nucleotide at a first interrogation position; and (c) a label; wherein each label of the set corresponds to the unique nucleotide.

[0319] In a further embodiment, and in accordance with any of the above, each of the monomers in a concatemer comprises a plurality of adaptors.

[0320] In a still further embodiment and in accordance with any of the above, at least one of the adaptors in a concatemer includes at least one Type IIs endonuclease recognition site.

[0321] In a still further embodiment and in accordance with any of the above, the steps of hybridizing a first anchor probe to the first anchor site; hybridizing a second anchor probe to a second anchor site; hybridizing at least a first sequencing probe to the first target domain; and ligating the anchor probes and the sequencing probe are repeated to identify a second nucleotide at a second detection position.

[0322] In a further embodiment and in accordance with any of the above, the second anchor probe includes a set of second anchor probes comprising at least three degenerate bases that hybridize to sequences outside the second anchor site.

[0323] In a still further embodiment and in accordance with any of the above, the second anchor probe comprises at least one terminus that is selectively activatable for ligation.

[0324] In a still further embodiment and in accordance with any of the above, the surface with a plurality of concatemers is a functionalized surface. In a still further embodiment, the surface is functionalized with functional moieties selected from a group including amines, silanes, and hydroxyls.

[0325] In a still further embodiment and in accordance with any of the above, the surface comprises a plurality of spatially distinct regions comprising the immobilized concatemers.

[0326] In a further embodiment and in accordance with any of the above, the concatemers are immobilized on the surface using capture probes.

[0327] In a still further embodiment and in accordance with any of the above, genomic nucleic acids are fragmented to form target sequences.

[0328] In a still further embodiment and in accordance with any of the above, the target sequence is a genomic nucleic acid sequence.

[0329] In a still further embodiment and in accordance with any of the above, the genomic nucleic acid sequences are human.

[0330] In one aspect and in accordance with any of the above, the present invention provides kits for use with sequencing templates that can include the sets of probes described herein. In general, kits of the invention can include anchor probe pairs, anchor probe pairs and an additional anchor probe adjacent to a target nucleic acid in a template, and sequencing probes for base determination at specific positions in the nucleic acid template. Such kits may further comprise adaptors for use in the generation of nucleic acid templates for use in the present methods.

[0331] In one aspect and in accordance with any of the above, the present invention provides a nucleic acid sequencing system that comprises 10 pools of labeled or tagged probes, sets of anchor probes comprising 4 or more probes of different sequences, sets of anchor probes with 3 or more degenerate bases, and ligase. In a further embodiment, the nucleic acid sequencing system will further comprise an agent for denaturing anchor probes, sequencing probes, and ligated sequencing and anchor probes from a nucleic acid template.

EXAMPLES

Example 1: Producing DNBs

[0332] The following are exemplary protocols for producing DNBs (also referred to herein as "amplicons") from nucleic acid templates of the invention comprising target nucleic acids interspersed with one or more adaptors. Single-stranded linear nucleic acid templates are first subjected to amplification with a phosphorylated 5' primer and a biotinylated 3' primer, resulting in a double-stranded linear nucleic acid templates tagged with biotin.

[0333] First, streptavidin magnetic beads were prepared by resuspending MagPrep-Streptavidin beads (Novagen Part. No. 70716-3) in 1x bead binding buffer (150 mM NaCl and 20 mM Tris, pH 7.5 in

nuclease free water) in nuclease-free microfuge tubes. The tubes were placed in a magnetic tube rack, the magnetic particles were allowed to clear, and the supernatant was removed and discarded. The beads were then washed twice in 800 µl 1x bead binding buffer, and resuspended in 80 µl 1x bead binding buffer. Amplified nucleic acid templates (also referred to herein as "library constructs") from the PCR reaction were brought up to 60 µl volume, and 20 µl 4x bead binding buffer was added to the tube. The nucleic acid templates were then added to the tubes containing the MagPrep beads, mixed gently, incubated at room temperature for 10 minutes and the MagPrep beads were allowed to clear. The supernatant was removed and discarded. The MagPrep beads (mixed with the amplified library constructs) were then washed twice in 800 µl 1x bead binding buffer. After washing, the MagPrep beads were resuspended in 80 µl 0.1N NaOH, mixed gently, incubated at room temperature and allowed to clear. The supernatant was removed and added to a fresh nuclease-free tube. 4 µl 3M sodium acetate (pH 5.2) was added to each supernatant and mixed gently.

[0334] Next, 420 µl of PBI buffer (supplied with QIAprep PCR Purification Kits) was added to each tube, the samples were mixed and then were applied to QIAprep Miniprep columns (Qiagen Part No. 28106) in 2 ml collection tubes and centrifuged for 1 minutes at 14,000 rpm. The flow through was discarded, and 0.75 ml PE buffer (supplied with QIAprep PCR Purification Kits) was added to each column, and the column was centrifuged for an additional 1 minute. Again the flow through was discarded. The column was transferred to a fresh tube and 50 µl of EB buffer (supplied with QIAprep PCR Purification Kits) was added. The columns were spun at 14,000 for 1 minute to elute the single-stranded nucleic acid templates. The quantity of each sample was then measured.

[0335] Circularization of single-stranded templates using CircLigase: First, 10 pmol of the single-stranded linear nucleic acid templates was transferred to a nuclease-free PCR tube. Nuclease free water was added to bring the reaction volume to 30 µl, and the samples were kept on ice. Next, 4 µl 10x CircLigase Reaction Buffer (Epicentre Part. No. CL4155K), 2 µl 1mM ATP, 2 µl 50 mM $MnCl_2$, and 2 µl CircLigase (100U/ µl) (collectively, 4x CircLigase Mix) were added to each tube, and the samples were incubated at 60°C for 5 minutes. Another 10 µl of 4x CircLigase Mix was added was added to each tube and the samples were incubated at 60° for 2 hours, 80°C for 20 minutes, then 4°C. The quantity of each sample was then measured.

[0336] Removal of residual linear DNA from CircLigase reactions by Exonuclease digestion. First, 30 µl of each CircLigase sample was added to a nuclease-free PCR tube, then 3 µl water, 4 µl 10x Exonuclease Reaction Buffer (New England Biolabs Part No. B0293S), 1.5 µl Exonuclease I (20 U/µl, New England Biolabs Part No. M0293L), and 1.5 µl Exonuclease III (100 U/µl, New England Biolabs Part No. M0206L) were added to each sample. The samples were incubated at 37°C for 45 minutes. Next, 75 mM EDTA, ph 8.0 was added to each sample and the samples were incubated at 85°C for 5 minutes, then brought down to 4°C. The samples were then transferred to clean nuclease-free tubes. Next, 500 µl of PN buffer (supplied with QIAprep PCR Purification Kits) was added to each tube, mixed and the samples were applied to QIAprep Miniprep columns (Qiagen Part No. 28106) in 2 ml collection tubes and centrifuged for 1 minute at 14,000 rpm. The flow through was discarded, and 0.75 ml PE buffer (supplied with QIAprep PCR Purification Kits) was added to each column, and the column was centrifuged for an additional 1 minute. Again the flow through was discarded. The column was transferred to a fresh tube and 40 µl of EB buffer (supplied with QIAprep PCR Purification Kits) was added. The columns were spun

at 14,000 for 1 minute to elute the single-stranded library constructs. The quantity of each sample was then measured.

[0337] Circle dependent replication for DNB production: The nucleic acid templates were subjected to circle dependent replication to create DNBs comprising concatamers of target nucleic acid and adaptor sequences. 40 fmol of exonucleoase-treated single-stranded circles were added to nuclease-free PCR strip tubes, and water was added to bring the final volume to 10.0. µl. Next, 10 µl of 2x Primer Mix (7 µl water, 2 µl 10x phi29 Reaction Buffer (New England Biolabs Part No. B0269S), and 1 µl primer (2 µM)) was added to each tube and the tubes were incubated at room temperature for 30 minutes. Next, 20 µl of phi 29 Mix (14 µl water, 2 µl 10x phi29 Reaction Buffer (New England Biolabs Part No. B0269S), 3.2 dNTP mix (2.5mM of each dATP, dCTP, dGTP and dTTP), and 0.8 µl phi29 DNA polymerase (10 U/µl, New England Biolabs Part No. M0269S)) was added to each tube. The tubes were then incubated at 30°C for 120 minutes. The tubes were then removed, and 75 mM EDTA, pH 8.0 was added to each sample. The quantity of circle dependent replication product was then measured.

[0338] Determining DNB quality: Once the quantity of the DNBs was determined, the quality of the DNBs was assessed by looking at color purity. The DNBs were suspended in amplicon dilution buffer (0.8x phi29 Reaction Buffer (New England Biolabs Part No. B0269S) and 10 mM EDTA, pH 8.0), and various dilutions were added into lanes of a flowslide and incubated at 30°C for 30 minutes. The flowslides were then washed with buffer and a probe solution containing four different random 12-mer probes labeled with Cy5, Texas Red, FITC or Cy3 was added to each lane. The flow slides were transferred to a hot block pre-heated to 30°C and incubated at 30°C for 30 minutes. The flow slides were then imaged using Imager 3.2.1.0 software. The quantity of circle dependent replication product was then measured.

Example 2: Single and Double c-PAL

[0339] Different lengths of fully degenerate second anchor probes were tested in a two anchor probe detection system. The combinations used were: 1) standard one anchor ligation using an anchor that binds to the adaptor adjacent to the target nucleic acid and a 9-mer sequencing probe, reading at position 4 from the adaptor 2) two anchor ligation using the same first anchor and a second anchor comprising a degenerate five-mer and a 9-mer sequencing probe, reading at position 9 from the adaptor; 3) two anchor ligation using the same first anchor and a second anchor comprising a degenerate six-mer and a 9-mer sequencing probe, reading at position 10 from the adaptor; and 4) two anchor ligation using the same first anchor and a second anchor comprising a degenerate eight-mer and a 9-mer sequencing probe, reading at position 12 from the adaptor. 1 µM of a first anchor probe and 6 µM of a degenerate second anchor probe were combined with T4 DNA ligase in a ligase reaction buffer and applied to the surface of the reaction slide for 30 minutes, after which time the unreacted probes and reagents were washed from the slide. A second reaction mix containing ligase and fluorescent probes of the type 5' Fl-NNNNNBNNN or 5' Fl-NNBNNNNNN 5' Fl-NNNBNNNNN 5' Fl-NNNNBNNNN was introduced. Fl represents one of four fluorophores, N represents any one of the four bases A, G, C, or T introduced at random, and B represents one of the four bases A, G, C, or T specifically associated with the fluorophore. After ligation for 1 hr the unreacted probes and reagents were washed from the slide and the fluorescence associated with each DNA target was assayed.

[0340] Figure 27 demonstrates signal intensities associated with the different length degenerate second anchor probes in the systems, with intensities decreasing with increased second anchor probe length. As seen in Figure 28, the fit scores for such intensities also decreased with the length of the degenerate second anchor, but still generated reasonable fit scores through the base 10 read.

[0341] The effect of time using the one anchor probe method and the two anchor probe method is illustrated in Figures 29 and 30. The standard anchor and degenerate five-mer were both used with a 9-mer sequencing probe to read positions 4 and 9 from the adaptor, respectively. Although the intensity levels differed more in the two anchor probe method, both the standard one anchor method and the two anchor probe methods at both times demonstrated comparable fit scores, each being over 0.8.

[0342] Effect of degenerate second anchor probe length on intensity and fit score: Different combinations of first and second anchor probes with varying second anchor probe length and composition were used to compare the effect of the degenerate anchor probe on signal intensity and fit score when used to identify a base 5' of the adaptor. Standard one anchor methods were compared to signal intensities and fit scores using two anchor probe methods with either partially degenerate probes having some region of complementarity to the adaptor, or fully degenerate second anchor probes. Degenerate second anchor probes of five-mers to nine-mers were used at one concentration, and two of these – the 6-mer and the seven-mer, were also tested at 4X concentration. Second anchor probes comprising two nucleotides of adaptor complementarity and different lengths of degenerate nucleotides at their 3' end were also tested at the first concentration. Each of the reactions utilized a same set of four sequencing probes for identification of the nucleotide present at the read position in the target nucleic acid.

[0343] The combinations used in the experiments are as follows:

Reaction 1: 1 μM of a 12 base first anchor probe
No second anchor probe
Read position: 2nt from the adaptor end

Reaction 2: 1 μM of a 12 base first anchor probe
20 μM of 5 degenerate base second anchor probe
Read position: 7nt from the adaptor end

Reaction 3: 1 μM of a 12 base first anchor probe
20 μM of a 6 degenerate base second anchor probe
Read position: 8nt from the adaptor end

Reaction 4: 1 μM of a 12 base first anchor probe
20 μM of a 7 degenerate base second anchor probe
Read position: 9nt from the adaptor end

Reaction 5: 1 μM of a 12 base first anchor probe
20 μM of an 8 degenerate base second anchor probe
Read position: 10nt from the adaptor end

Reaction 6: 1 μM of a 12 base first anchor probe
20 μM of a 9 degenerate base second anchor probe
Read position: 11nt from the adaptor end

Reaction 7: 1 μM of a 12 base first anchor probe
80 μM of a 6 degenerate base second anchor probe

Read position: 8nt from the adaptor end

Reaction 8: 1 μM of a 12 base first anchor probe
80 μM of a 7 degenerate base second anchor probe
Read position: 9nt from the adaptor end

Reaction 9: 1 μM of a 12 base first anchor probe
20 μM of a 6nt second anchor probe (4 degenerate bases - 2 known bases)
Read position: 6nt from the adaptor end

Reaction 10: 1 μM of a 12 base first anchor probe
20 μM of a 7nt second anchor probe (5 degenerate bases - 2 known bases)
Read position: 7nt from the adaptor end

Reaction 11: 1 μM of a 12 base first anchor probe
20 μM of an 8nt second anchor probe (6 degenerate bases - 2 known bases)
Read position: 8nt from the adaptor end

[0344] Figures 31 and 32 illustrate the results of the different combinations of anchor probes and sequencing probes. The length of the degenerate second anchor probe was shown to be best using a six-mer, whether it was completely degenerate or partially degenerate. The signal intensities using a fully degenerate six-mer at a higher concentration showed signal intensities similar to that of the partially degenerate six-mer (Figure 31). All data had fairly good fit scores (see Figure 32) except Reaction 6 using the longest of the second anchors, which also displayed the lowest intensity scores of the reactions performed (Figure 31).

[0345] Effect of first anchor probe length on intensity and fit score: Different combinations of first and second anchor probes with varying first anchor probe length were used to compare the effect of the first anchor probe length on signal intensity and fit score when used to identify a base 3' of the adaptor. Standard one anchor methods were compared to signal intensities and fit scores using two anchor probe methods with either partially degenerate probes having some region of complementarity to the adaptor, or fully degenerate second anchor probes. Each of the reactions utilized a same set of four sequencing probes for identification of the nucleotide present at the read position in the target nucleic acid. The combinations used in the experiment are as follows:

Reaction 1: 1 μM of a 12 base first anchor probe
No second anchor probe
Read position: 5nt from the adaptor end

Reaction 2: 1 μM of a 12 base first anchor probe
20 μM of 5 degenerate base second anchor probe
Read position: 10nt from the adaptor end

Reaction 3: 1 μM of a 10 base first anchor probe
20 μM of a 7nt second anchor probe (5 degenerate bases - 2 known bases)

Read position: 10nt from the adaptor end

Reaction 4: 1 µM of a 13 base first anchor probe
20 µM of a 7 degenerate base second anchor probe
Read position: 12nt from the adaptor end

Reaction 5: 1 µM of a 12 base first anchor probe
20 µM of an 7 degenerate base second anchor probe
Read position: 12nt from the adaptor end

Reaction 6: 1 µM of a 11 base first anchor probe
20 µM of a 7 degenerate base second anchor probe
Read position: 12nt from the adaptor end

Reaction 7: 1 µM of a 10 base first anchor probe
20 µM of a 7 degenerate base second anchor probe
Read position: 12nt from the adaptor end

Reaction 8: 1 µM of a 9 base first anchor probe
80 µM of a 7 degenerate base second anchor probe
Read position: 12nt from the adaptor end

[0346] The signal intensity (Figure 33) and fit scores (Figure 34) observed show an optimum intensity resulting from use of the longer first anchor probes, which in part may be due to the greater meting temperature the longer probes provide to the combined anchor probe.

[0347] Effect of kinase incubations on intensity and fit score using two anchor primer methods: The reactions as described above were performed at different temperatures using 1 µM of a 10 base first anchor probe, 20 µM of a 7-mer second anchor probe, and sequencing probe with the structure Fluor-NNNNBNNNN to read position 10 from the adaptor in the presence of a kinase at 1Unit/ml for a period of three days. A reaction with a 15-mer first anchor and the sequencing probe served as a positive control. Results are as shown in Figures 35 and 36. Although the kinase did have an effect on signal intensities as compared to the control, the range did not change from 4°C to 37°C, and fit scores remained equivalent with the control. The temperature at which the kinase incubation did have an impact is 42°C, which also displayed a poor fit with the data.

[0348] The minimum time needed to kinase was then examined using the same probes and conditions as described above. As shown in Figures 37 and 38, kinase incubation of five minutes or above resulting in effectively equivalent signal intensities and fit score.

[0349] 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. Unless otherwise clear from the context or expressly stated, any concentration values provided herein are generally given in terms of admixture values or percentages without regard to any conversion that occurs upon or following addition of the particular component of the mixture. To the extent not already expressly incorporated herein, all published references and patent documents referred to in this disclosure are incorporated herein by reference in their entirety for all purposes. Changes in detail or structure may be made without departing from the basic elements of the present technology as defined in the following claims.

CLAIMS

What is claimed:

1. A method of determining the identification of a first nucleotide at a detection position of a target sequence comprising a plurality of detection positions, said method comprising:

(a) providing a surface comprising a plurality of concatemers, wherein each concatemer comprises a plurality of monomers and each monomer comprises:

i) a first target domain of said target sequence comprising a first set of target detection positions;

ii) at least a first adaptor comprising:

1) a first anchor site; and

2) a second adjacent anchor site;

b) hybridizing a first anchor probe to said first anchor site;

c) hybridizing a second anchor probe to said second anchor site, wherein said second anchor probe also hybridizes to sequences outside said second anchor site;

d) hybridizing at least a first sequencing probe to said first target domain, wherein said first sequencing probe comprises:

i) a first probe domain complementary to said target domain;

ii) a unique nucleotide at a first interrogation position; and

iii) a label;

under conditions wherein if said unique nucleotide is complementary to said first nucleotide, said sequencing probe hybridizes to said concatemer; and

e) ligating said anchor probes and said sequencing probe;

f) identifying said first nucleotide.

2. A method according to claim 1 wherein a set of sequencing probes are contacted with said surface, each sequencing probe comprising:

a) a first probe domain complementary to said target domain;

b) a unique nucleotide at a first interrogation position; and

c) a label;

wherein each label of said set corresponds to said unique nucleotide.

3. A method according to claim 1 wherein each of said monomers comprises a plurality of adaptors.

4. A method according to claim 3 wherein at least one of said adaptors comprises at least one Type IIs endonuclease recognition site.

5. A method according to claim 1 wherein steps b) – e) are repeated to identify a second nucleotide at a second detection position.

6. A method according to claim 1, wherein said second anchor probe comprises a set of second anchor probes comprising at least three degenerate bases that hybridizes to said sequences outside said second anchor site.

7. A method according to claim 1, wherein said second anchor probe comprises at least one terminus that is selectively activatable for ligation.

8. A method according to claim 1 wherein said surface is functionalized.

9. A method according to claim 8 wherein said functionalized surface comprises functional moieties selected from the group consisting of amines, silanes, and hydroxyls.

10. A method according to claim 1 wherein said surface comprises a plurality of spatially distinct regions comprising said immobilized concatemers.

11. A method according to claim 1 wherein said concatemers are immobilized on said surface using capture probes.

12. A method according to claim 1 further comprising fragmenting genomic nucleic acid to form target sequences.

13. A method according to claim 1 wherein said target sequence is a genomic nucleic acid sequence.

14. A substrate according to claim 13 wherein said genomic nucleic acid sequences are human.