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

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
22 July 2010 (22.07.2010)

PCT

(10) International Publication Number
WO 2010/083046 A2

(51) International Patent Classification:
C12Q 1/68 (2006.01) *G01N 33/53* (2006.01)
C12N 15/11 (2006.01)

(21) International Application Number:
PCT/US2010/000102

(22) International Filing Date:
15 January 2010 (15.01.2010)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
61/205,397 15 January 2009 (15.01.2009) US

(71) Applicant *(for all designated States except US)*: **THE SALK INSTITUTE FOR BIOLOGICAL STUDIES** [US/US]; 10010 N. Torrey Pines Road, La Jolla, CA 92037 (US).

(72) Inventors; and
(75) Inventors/Applicants *(for US only)*: **YEO, Gene** [SG/US]; 3855 Nobel Drive #2210, San Diego, CA 92122 (US). **SCOLNICK, Jonathan** [US/SG]; 156 Prince Charles Crescent, Tower A #02-11, Singapore 159015 (SG). **GAGE, Fred, H.** [US/US]; 1533 Virginia Way, La Jolla, CA 92037 (US).

(74) Agents: **QUINE, Jonathan, Alan** et al.; Quine Intellectual Property Law Group, P.C, P.O.box 458, Alameda, CA 94501 (US).

(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IS, JP, KE, KG, KM, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PE, PG, PH, PL, PT, RO, RS, RU, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(84) Designated States *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, GH, GM, KE, LS, MW, MZ, NA, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Published:

— *without international search report and to be republished upon receipt of that report (Rule 48.2(g))*

(54) Title: METHODS FOR USING NEXT GENERATION SEQUENCING TO IDENTIFY 5-METHYL CYTOSINES IN THE GENOME

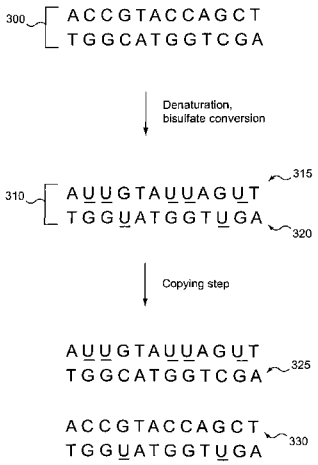


Fig. 3

(57) Abstract: Provided are methods for producing a set of discriminator probes that selectively hybridize to unmethylated sequences in a population of nucleic acids, which sequences have been treated with a methylation state conversion reagent. Also provided are methods of using the discriminator probes to separate methylated subsequences from unmethylated subsequences in a population of nucleic acids. Compositions comprising said discriminator probes are provided. In addition, compositions comprising said discriminator probes hybridized to unmethylated sequences in a population of nucleic acid, which unmethylated sequences have been treated with a methylation state conversion reagent are provided.

WO 2010/083046 A2

METHODS FOR USING NEXT GENERATION SEQUENCING TO IDENTIFY 5-METHYL CYTOSINES IN THE GENOME

CROSS-REFERENCE TO RELATED APPLICATIONS

[0001] This application claims priority to and benefit of U. S. Provisional Patent Application 61/205,397, entitled, "Methods for Using Next Generation Sequencing to Identify 5-methyl Cytosines in the Genome", by Yeo, Scolnick, and Gage, filed January 15, 2009, the disclosure of which is incorporated herein in its entirety for all purposes.

FIELD OF THE INVENTION

[0002] The invention relates to molecular methods that can be used in highly parallel analyses to determine the methylation pattern of a nucleic acid, e.g., a genomic DNA.

BACKGROUND OF THE INVENTION

[0003] DNA methylation is a well-characterized, heritable epigenetic modification that is essential in mammals (Li, et al. (1992) "Targeted mutation of the DNA methyltransferase gene results in embryonic lethality." *Cell* **69:** 915–926). The methylation patterns present in a mammalian genome can affect a wide variety of biological processes, including, e.g., embryonic development, transcription, chromatin structure, X chromosome inactivation, genomic imprinting, drug activity, and chromosome stability. A growing number of human diseases have been found to be associated with aberrant DNA methylation, e.g., cancer, lupus, Alzheimer's, and others (Iacobuzio-Donahue (2008) "Epigenetic Changes in Cancer." *Annu Rev Pathol* doi:10.1146/annurev.pathol.3.121806.151442; Robertson (2004) "DNA methylation and human disease." *Nature Rev Genet* **6:** 597-610; Toyota, et al. (2002) "DNA methylation changes in gastrointestinal disease." *J Gastroenterol* **37:** 97-101; Mastubayashi, et al. (2005) "Age- and disease-related methylation of multiple genes in non-neoplastic duodenum and in duodenal juice." *Clin Cancer Res* **11:** 573-583; Richardson (2003) "DNA methylation and autoimmune disease" *Clin Immun* **109:** 72-79). Genomic methylation profiles can therefore be of particularly advantageous use in determining an individual's predisposition to a disease, in diagnosing a disease state, in predicting a patient's prognosis, and/or in determining a therapeutic regimen.

[0004] Methylation specific PCR (MSP) is one strategy that can be used to analyze methylation patterns at specific loci in a genome (Herman, et al. (1996) "Methylation-specific PCR: a novel PCR assay for methylation status of CpG islands." *Proc Natl Acad Sci USA* **93:** 9821-9826; Cottrell, et al. (2004) "A real-time PCR assay for DNA-methylation using methylation-specific blockers." *Nucleic Acids Res* **32:** e10; and Thomassin, et al. (2004) "MethylQuant: a sensitive method for quantifying methylation of specific cytosines within the genome." *Nucleic Acids Research* **32:** e168). However, this technique is both time consuming and labor-intensive, and PCR is not easily scalable for the efficient high-throughput interrogation of the complete methylation profile of, e.g., a large genome.

[0005] Methylation-specific oligonucleotide (MSO) microarray analysis is one currently available method for the highly parallel detection of methylation pattern variations in, e.g., a genomic DNA. In MSO, oligonucleotides that correspond to methylated and unmethylated alleles of, e.g., a region of interest in a genome, are affixed to a solid support and used to probe, e.g., products amplified from sodium bisulfite-treated DNA. (Gitan, et al. (2002) "Methylation-Specific Oligonucleotide Microarray: A New Potential for High-Throughput Methylation Analysis." *Genome Res* doi: 10.1101/gr.202801; Adorjan, et al. (2002) "Tumour class prediction and discovery by microarray-based DNA methylation analysis." *Nucleic Acids Res* doi: 10.1093/nar/30.5.e21). The probes discriminate methylated and unmethylated cytosines at specific nucleotide positions, and quantitative differences in hybridization are determined by fluorescence analysis. However, because this microarray technique is restricted to detecting the methylation patterns of specific sequences, it can be used more effectively in the validation and/or high-resolution analysis of candidate regions. Furthermore, closely spaced regions of interest in a genome may not be amenable to MSO microarray analysis if the gene in question is heterogeneously methylated.

[0006] What are needed in the art are molecular tools and cost-effective, high-throughput methods for the detection and characterization of DNA methylation patterns present in, e.g., a genome. The invention described herein fulfills these and other needs, as will be apparent upon review of the following.

SUMMARY OF THE INVENTION

[0007] The present invention provides methods and related compositions useful for distinguishing and separating unmethylated sequences from methylated sequences in a nucleic acid sample that comprises both methylated and unmethylated subsequences. In the methods, unmethylated subsequences in the sample, which subsequences have been converted as a result of treatment with a methylation state conversion reagent, selectively hybridize to a set of discriminator probes and can optionally be subtracted from the methylated subsequences. This separation enriches the population for methylated sequences, which can then be further characterized, e.g., via sequencing, e.g., using an automated high-throughput sequencing system. The methods and compositions provided by the invention can advantageously permit high throughput methylation profiling of, e.g., a large mammalian genome. Such profiles would otherwise be difficult to obtain using current methods.

[0008] Thus, in a first aspect, the invention provides methods of distinguishing unmethylated subsequences from methylated subsequences in a nucleic acid sample. The methods include providing the nucleic acid sample, fragmenting the nucleic acid sample to produce fragments, and treating the fragments with a methylation state conversion reagent. This treatment produces treated nucleic acids that comprise a subset of converted nucleic acids and a subset of unconverted nucleic acids, wherein the converted nucleic acids correspond to the unmethylated subsequences in the nucleic acid sample and wherein the unconverted nucleic acids correspond to methylated subsequences in the nucleic acid sample. The treated nucleic acids are then hybridized to a set of discriminator probes that selectively hybridize to the converted nucleic acids to produce hybridized nucleic acids, thereby distinguishing unmethylated subsequences from methylated subsequences in a nucleic acid sample.

[0009] Optionally, both the population of nucleic acids and the discriminator probes can be derived from a first source of nucleic acids, e.g., a genomic DNA. The nucleic acid sample can be fragmented using any of a number of methods, including, e.g., enzymatic digestion, sonication, mechanical shearing, electrochemical cleavage, and/or nebulization. In preferred embodiments, the methylation state conversion reagent with which the nucleic acid sample is treated is sodium bisulfite.

[0010] These methods optionally comprise a further step wherein the hybridized nucleic acids are separated from the unconverted nucleic acids. The separation step can be performed using any of a number of strategies. For example, separating the hybridized nucleic acids from the unconverted nucleic acids can optionally comprise electrophoresis. In other embodiments, the converted nucleic acids can hybridize to tagged discriminator probes, e.g., probes that comprise any one or more of the following moieties: a ligand, a fluorescent label, a blocking group, a phosphorylated nucleotide, a biotinylated nucleotide, a methylated nucleotide, a uracil, a sequence capable of forming hairpin secondary structure, an oligonucleotide hybridization site, a restriction site, a DNA promoter, a protein binding sequence, a sample or library identification sequence, a thiol linker, a phosphorothioated nucleotide, an amine reactive nucleotide, and a cis regulatory sequence. In such an embodiment, the tagged hybridized nucleic acids can be separated from the unconverted nucleic acids via affinity purification.

[0011] The methods can optionally include sequencing the unconverted nucleic acids, e.g., using an automated high-throughput sequencing system, and comparing sequences of the unconverted fragments to sequences of the nucleic acid sample to identify the methylated subsequences in the sample.

[0012] The invention also provides compositions that include a population of unconverted nucleic acid acids that have been produced using the methods described above.

[0013] In a related aspect, the invention provides compositions comprising a set of discriminator probes that selectively hybridize to converted nucleic acids in a population of nucleic acids that has been treated with a methylation state conversion reagent and a population of nucleic acids that comprises a subset of methylated nucleic acids and a subset of unmethylated nucleic acids, which population has been treated with the conversion reagent to produce a subset of converted nucleic acids and a subset of unconverted nucleic acids, wherein the converted nucleic acids correspond to the unmethylated nucleic acids, and wherein the unconverted nucleic acids correspond to the methylated nucleic acids. In these compositions, the converted nucleic acids are hybridized to the set of discriminator probes. Optionally, the discriminator probes and the population of nucleic acids in the compositions are both derived from a first source of nucleic acids, e.g., a genomic DNA. Optionally, the methylation state conversion reagent with which the population of nucleic acids has been treated is sodium bisulfite.

[0014] The invention also provides a composition comprising a set of probes capable of distinguishing between unmethylated nucleic acid sequences and methylated nucleic acids sequences.

[0015] In another aspect, the invention provides methods of producing a set of discriminator probes that selectively hybridize to converted sequences in a nucleic acid sample that has been treated with a methylation state conversion reagent. The methods include providing at least one nucleic acid that corresponds to a sequence present in the nucleic acid sample. In the methods, the nucleic acid is fragmented, e.g., optionally using any of the methods described above, and the resulting nucleic acid fragments are amplified, e.g., via PCR and/or primer extension, to produce a population of unmethylated nucleic acids. The population of unmethylated nucleic acids is then treated with the methylation conversion reagent, e.g., sodium bisulfite, to produce converted nucleic acids. The converted nucleic acids are then copied to produce the set of discriminator probes.

[0016] Copying the converted fragments can include, e.g., annealing tagged DNA primers to 3' ends of the converted fragments and extending the tagged primers with a polymerase. Optionally, the tagged DNA primers can comprise any of the moieties described above.

[0017] Optionally, the at least one nucleic acid from which the discriminator probes are produced and the nucleic acid sample can both be derived from a first source of nucleic acids, e.g., a genomic DNA

[0018] One of skill in the art will appreciate that the methods and compositions provided by the invention can be advantageously used in alone or in combination with one another.

[0019] Kits are also a feature of the invention. The present invention provides kits that include useful reagents, e.g., tagged DNA primers, affinity columns, and/or one or more enzyme and/or reagent that are used in the methods, e.g., a DNA polymerase, bisulfite, etc. Such reagents are most preferably packaged in a fashion to enable their use. The kits of the invention optionally include additional reagents, such as a control target nucleic acids, buffer solutions and/or salt solutions, including, e.g., divalent metal ions, i.e., Mg^{++}, Mn^{++} and/or Fe^{++}, nucleic acid adapter tags, e.g., to prepare methylated nucleic acid fragments for sequencing, e.g., using a currently available or future automated high-throughput

sequencing system. Such kits also typically include a container to hold the kit components, instructions for use of the compositions, e.g., to practice the methods.

DEFINITIONS

[0020] Before describing the present invention in detail, it is to be understood that this invention is not limited to particular devices or biological systems, which can, of course, vary. It is also to be understood that the terminology used herein is for the purpose of describing particular embodiments only, and is not intended to be limiting. As used in this specification and the appended claims, the singular forms "a", "an" and "the" include plural referents unless the content clearly dictates otherwise. Thus, for example, reference to "a discriminator probe" includes a combination of two or more discriminator probes; reference to "nucleic acid fragments" includes mixtures of nucleic acid fragments, and the like.

[0021] 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 the invention pertains. Although any methods and materials similar or equivalent to those described herein can be used in the practice for testing of the present invention, the preferred materials and methods are described herein. In describing and claiming the present invention, the following terminology will be used in accordance with the definitions set out below.

[0022] Copying: As used herein, "copying" refers to the process of replicating a nucleic acid molecule to generate a new nucleic acid that comprises a sequence complementary to that of the original template molecule.

[0023] Derived from: As used herein, "derived from" is used to refer to the original source organism, tissue, cells, etc. from which, e.g., a nucleic acid sample and/or a set discriminator probes to be used with the methods of the invention, was obtained. For example, nucleic acid samples and/or set of discriminator probes can optionally be derived from, e.g., a cell line or a eukaryotic organism, including, but not limited to, mammals, nematodes, insects, etc.

[0024] Linker: As used herein, a linker is a single-stranded nucleic acid of about 2-20 nucleotides in length (or longer) that can be attached to a single stranded nucleic acid, such as a denatured DNA, via ligation or by extending the linker, e.g., with a polymerase.

A nucleic acid linker can include any one or more of an oligonucleotide hybridization site, a restriction site, a DNA promoter, a protein binding site, a sample or library identification sequence, a thiol linker, a phosphorothioated nucleotide, an amine-reactive nucleotide, a cis regulatory sequence, modified nucleotide or nucleotide analog, and/or the like.

[0025] Methylation state conversion reagent: As used herein, a "methylation state conversion reagent" is a reagent that introduces specific changes to a nucleic acid sequence based on the methylation status of particular nucleotide residues in the sequence. For example, sodium bisulfate, a methylation state conversion reagent used in preferred embodiments of the invention, preferentially deaminates unmethylated cytosine residues to uracils, leaving 5-methylcytosine residues unreacted. Preferably, treating a nucleic acid sample with a methylation state conversion reagent will result in the complete stoichiometric conversion of, e.g., unmethylated cytosine residues, into, e.g., uracil residues. Nucleic acids that comprise residues that have been changed as a result of treatment with a methylation state conversion reagent are referred to herein as "converted nucleic acids", and nucleotide sequences that have been changed as a result of such treatment are referred to herein as "converted sequences". Therefore, a converted sequence is produced by the reaction of unmethylated residues with the methylation state conversion reagent. An unconverted sequence comprises methylated residues, which are not susceptible to converson by the reagent. For example, a nucleic acid comprising methylated nucleotides can undergo treatment with a methylation state conversion reagent. However, because the methylated nucleotides are not affected by the treatment, the treated nucleic acid is not a converted nucleic acid.

[0026] Discriminator probe: As used herein, a "discriminator probe" is a nucleic acid probe that selectively hybridizes to unmethylated sequences in a population of nucleic acids that has been treated with a methylation state conversion reagent. Discriminator probes comprise sequences that are complementary to "converted sequences", e.g., sequences that have been changed as a result of treatment with a methylation specific conversion reagent, and not to "unconverted sequences", e.g., sequences that comprise methylated residues, i.e., residues that are not susceptible to the reagent. In other words, the sequence specificity of the discriminator probes precludes them from forming duplexes with nucleic acids that comprise "unconverted" (i.e., methylated) sequences. A set of discriminator probes can comprise at least 2 probes and can include up to as many probes as

are necessary to interrogate the entire sequence of a genome, e.g., a large mammalian genome, e.g., a human genome.

[0027] Tags: As used herein, a "tag" refers to a moiety linked to a molecule of interest that can be used as a molecular label to detect the molecule of interest in a population and/or as a tool by which to separate the molecule of interest from the population. For example, tags can be hybridized to the ends of the nucleic acid fragments and extended with a polymerase to produce tagged fragments or ligated to the ends of the nucleic acid fragments with a ligase. Tags can comprise any one or more moieties that include, e.g., a ligand, a fluorescent label, a blocking group, a phosphorylated nucleotide, a nucleotide analog, a fluorinated nucleotide, a nucleotide comprising a heavy atom, a biotinylated nucleotide, a methylated nucleotide, a uracil, a sequence capable of forming hairpin secondary structure, an oligonucleotide hybridization site, a restriction site, a DNA promoter, a protein binding site, a sample or library identification sequence, a thiol linker, a phosphorothioated nucleotide, an amine-reactive nucleotide, and/or a cis regulatory sequence.

[0028] Treatment: As used herein, "treatment" refers to a the exposure of, e.g., a nucleic acid to a methylation state conversion reagent. Preferably, the exposure of the nucleic acid to the treatment will result in, e.g., the stoichiometric conversion of a particular set of unmethylated nucleotides into a different set of nucleotides. For example, treatment of a nucleic acid with, e.g., sodium bisulfite, results in the conversion of unmethylated cytosines present in the nucleic acid into uracils. However, not all treated nucleic acids become converted nucleic acids. For example, sequences that comprise only methylated residues, i.e., residues that are not susceptible to the methylation state conversion reagent, are not converted by the reagent. Treatment of methylated nucleic produces unconverted nucleic acids.

BRIEF DESCRIPTION OF THE DRAWINGS

[0029] **Figure 1** is schematic depiction of how methods of the invention can be used to separate unmethylated subsequences of a nucleic acid from methylated subsequences.

[0030] **Figure 2** provides a schematic depiction of how methods provided by the invention can be used to produce a set of discriminator probes.

[0031] **Figure 3** provides a schematic that illustrates how discriminator probes selectively hybridize to unmethylated sequences that have been treated with a methylation state conversion reagent.

DETAILED DESCRIPTION

OVERVIEW

[0032] The impact of DNA methylation on genome-wide gene regulation, development, differentiation, and disease is largely uncharacterized. Identifying the methylation pattern present at a particular genomic locus in, e.g., a cell or tissue, e.g., at a specific developmental stage, in a disease state, or in response to a particular environmental stimulus, can provide insight into a variety of regulatory pathways, and can thus be advantageously useful in gaining insight into the study of mammalian embryogenesis, tissue morphogenesis, drug discovery, and disease. The methods and compositions provided herein can be used in combination for reliable, cost-effective identification of methylated sequences in, e.g., a genomic DNA.

[0033] The methods of the invention provide several advantages over currently available high-throughput technologies. For example, methylation-specific PCR (MSP) has been used to analyze methylation patterns at specific loci in a genome (Herman, et al. (1996) "Methylation-specific PCR: a novel PCR assay for methylation status of CpG islands." *Proc Natl Acad Sci USA* **93:** 9821-9826; Cottrell, et al. (2004) "A real-time PCR assay for DNA-methylation using methylation-specific blockers." *Nucleic Acids Res* **32:** e10; and Thomassin, et al. (2004) "MethylQuant: a sensitive method for quantifying methylation of specific cytosines within the genome." *Nucleic Acids Research* **32:** e168). However, this technique is both time consuming and labor-intensive, and PCR is not easily scalable for the efficient high-throughput interrogation of the complete methylation profile of, e.g., a large genome.

[0034] Methylation-specific oligonucleotide (MSO) microarray analysis is one currently available method for the highly parallel detection of methylation pattern variations in, e.g., a genomic DNA. In MSO, oligonucleotides that correspond to methylated and unmethylated alleles of, e.g., a region of interest in a genome, are affixed to a solid support and used to probe, e.g., products amplified from sodium bisulfite-treated DNA. (Gitan, et al. (2002) "Methylation-Specific Oligonucleotide Microarray: A New Potential for High-Throughput Methylation Analysis." *Genome Res* doi: 10.1101/gr.202801; Adorjan, et al.

(2002) "Tumour class prediction and discovery by microarray-based DNA methylation analysis." *Nucleic Acids Res* doi: 10.1093/nar/30.5.e21). The probes discriminate methylated and unmethylated cytosines at specific nucleotide positions, and quantitative differences in hybridization are determined by fluorescence analysis. However, because this microarray technique is restricted to detecting the methylation patterns of specific sequences, it can be used more effectively in the validation and/or high-resolution analysis of candidate regions. Furthermore, closely spaced regions of interest in a genome may not be amenable to MSO microarray analysis if the gene in question is heterogeneously methylated.

[0035] Restriction landmark genome scanning (RLGS) is another method for the analysis of global genomic methylation patterns. RLGS provides a quantitative assessment of thousands of CpG islands in a single gel without prior knowledge of gene sequence (Ando and Hayashizaki (2007) "Restriction Landmark Genome Scanning." *Nature Protocols* **1:** 1, 2774–2783). This method entails digesting genomic DNA with a "landmark enzyme", e.g., a restriction enzyme that does not cut methylated DNA (such as Not I or Asc I), radiolabeling the cleaved ends, digesting the radiolabeled fragments with a second restriction enzyme, and then electrophoresing the twice-digested genomic DNA fragments through a narrow, tube-shaped agarose gel. The DNA in the tube gel is then digested by a third, more frequently cutting restriction enzyme and electrophoresed, in a direction perpendicular to the first separation, through a 5% non-denaturing polyacrylamide gel, which is then autoradiographed.

[0036] When comparing two RLGS profiles, the absence of a radiolabeled fragment or "spot" at a particular position in an autoradiograph can indicate that a subsequence present in a DNA fragment that typically occupies that position on an autoradiograph has been methylated, and, therefore, cannot be cut by, e.g., Not I or Asc I. In contrast, the appearance of a radiolabeled fragment (or "spot") at a particular position in an autoradiograph can indicate the demethylation of, e.g., a locus in a genomic DNA, that is typically methylated. This approach has many drawbacks, from incomplete restriction-enzyme cutting to limitation of the regions that can be studied. However, the most significant caveat is the difficulty of cloning individual spots to determine the sequence of, e.g., the methylated genomic DNA fragment of interest. Cloning and sequencing a

fragment of interest can entail, e.g., Southern blotting or digital karyotyping, which are labor-intensive and time-consuming.

[0037] The detailed description is organized to first elaborate the methods and compositions provided by the invention for identifying methylated subsequences in a nucleic acids. Next, details regarding DNA methylation are described. Details regarding sequencing reactions and high-throughput sequencing systems are then provided. Kits, systems, and broadly applicable molecular biological techniques that can be used to perform any of the methods are described thereafter.

METHODS AND COMPOSITIONS FOR IDENTIFYING METHYLATED SUBSEQUENCES IN A NUCLEIC ACID

[0038] As described above, the methods provided herein can be used to identify the methylation pattern present at, e.g., a particular genomic locus, at a particular developmental stage, after exposure to a particular environmental stimulus, e.g., in a patient with a disease, etc. To perform the methods, at least one nucleic acid that comprises both methylated and unmethylated sequences, e.g., genomic DNA **100**, is provided (see **Figure 1**). Genomic DNA **100** is fragmented, using any one or more of the methods described hereinbelow, producing double-stranded nucleic acid fragments **110**. In preferred embodiments, fragments **110** are optionally, e.g., about 10 to 50 base pairs long, about 50 to 75 base pairs long, or about 75 to 100 base pairs long. Preferred fragments are about 50 base pairs long. Fragments **110** are then optionally denatured, if necessary, e.g., using any of a variety of methods known in the art, to produce single-stranded nucleic acids **115**, which are treated with a methylation state conversion reagent, e.g., a reagent that introduces specific changes into a nucleic acid sequence, which changes depend on the methylation status of individual residues. Treating single-stranded nucleic acids **115** with a methylation state conversion reagent produces treated nucleic acids **120**. Treatment leaves nucleic acids comprising methylated residues in an unconverted state, whereas nucleic acids comprising unmethylated residues react with the conversion reagent to produce converted sequences. Therefore, treated nucleic acids **120** comprise a subset of unconverted nucleic acids **155**, which correspond to methylated subsequences in genomic DNA **100**, and a subset of converted nucleic acids **150**, which correspond to unmethylated subsequences in genomic DNA **100**. Unconverted nucleic acids **155** comprise sequences that still have methylated

residues (i.e., unconverted residues) after treatment. However, these sequences may also contain some converted residues.

[0039] For example, in preferred embodiments of the methods described herein, denatured nucleic acid fragments, e.g., single-stranded nucleic acids **115**, are treated with the methylation state conversion reagent sodium bisulfite. Under certain conditions, treatment with sodium bisulfite, stoichiometrically converts cytosine (C) residues present in, e.g., single-stranded nucleic acids **115,** to uracil (U), but leave 5-methylcytosines unchanged. Resulting treated nucleic acids **120** are then hybridized to set of discriminator probes **130**.

[0040] Discriminator probes **240** in set **130** specifically hybridize to converted subsequences **150** present in treated nucleic acids **120**. In other words, discriminator probes form duplexes with nucleic acids that comprise sequences that contained only unmethylated cytosine residues (i.e., no methylated cytosine residues) prior to treatment with, e.g., sodium bisulfite. The probes are designed to exploit the sequence differences that distinguish unconverted (i.e., fully methylated) nucleic acids **155** from converted (i.e., unmethylated) nucleic acids **150**, e.g., sequence differences resulting from, e.g., the conversion of unmethylated cytosines to uracil via sodium bisulfite treatment, as detailed below. Following hybridization, the duplexes formed by discriminator probes **240** and converted nucleic acids **150** are removed, leaving population of methylated nucleic acids **165**, which correspond to unconverted nucleic acids **155**.

[0041] Compositions comprising a set of discriminator probes are provided by the invention, as are methods of producing discriminator probes. Producing a set of discriminator probes entails providing a nucleic acid, e.g., nucleic acid **200** (see **Figure 2**). Typically, the nucleic acid from which the discriminator probes are derived comprises at least one subsequence that corresponds to a subsequence present in the nucleic acid or population of nucleic acids that will be interrogated with the discriminator probes. For example, subsequence **205** in nucleic acid **200** corresponds to subsequence **105** present in genomic DNA **100** (see **Figure 1**). In general, the subsequence present in the nucleic acid from which the probes are derived, e.g., nucleic acid **200**, will comprise both methylated and unmethylated nucleotides.

[0042] Methods of producing discriminator probes from nucleic acid **200** include fragmenting nucleic acid **200**, e.g., using any one or more of the methods described hereinbelow. The resulting fragments are then amplified via, e.g., PCR or primer extension, to produce amplified fragments **227**. Unlike the template fragments from which they are amplified, fragments **227** do not comprise any methylated sequences. Amplified fragments **227** are then optionally denatured to produce single stranded nucleic acids **228**, which are treated with a methylation state conversion reagent, e.g., sodium bisulfite. Treatment of single stranded nucleic acids **228** with a methylation state conversion produces converted nucleic acids **230**, in which all cytosines have been stoichiometrically deaminated into uracils (see **Figure 2**). Converted nucleic acids **230** are then copied, e.g., using PCR, primer extension or the like, to produce discriminator probes **240**, which are complementary to converted nucleic acids **230**. In preferred embodiments, tags **241** are attached to discriminator probes **240**, using methods well known to those of skill in the art. Tags permit the separation of unmethylated sequences in a population of nucleic acids from methylated sequences, e.g., via affinity purification or other separation methods, as will be described in further detail below. Preferably, discriminator probes **240** will bind fully converted nucleic acids having no methylated residues (e.g., converted nucleic acids **150** in **Figure 1)** and will not bind unconverted sequences, e.g., those having at least one methylated residue (e.g., unconverted nucleic acids **155** in **Figure 1**).

[0043] Copying step **234** in **Figure 2** generates probes whose sequences are complementary only to unmethylated sequences that have been converted with a methylation state conversion reagent, e.g., sodium bisulfite. As shown in **Figure 3**, sodium bisulfite conversion of, e.g., fragment **300**, generates converted strands **315** and **320**, which are no longer complementary. Copying non-complementary strands **315** and **320** yields copied strands **325** and **330** whose sequences, under selected conditions, do not hybridize to either strand in fragment **300,** but can hybridize to converted unmethylated sequences, e.g., sequences present in converted fragment **310**, e.g., sequences that can be present in population **120** (see **Figure 1** and corresponding description). The sequences to which, e.g., strands **325** and **330** can hybridize, can then be subtracted from a population of nucleic acids that comprise methylated and unmethylated subsequences. The remaining nucleic acids, i.e., those that have not been subtracted, will comprise methylated subsequences.

[0044] Composition **160** (see **Figure 1**), which includes unhybridized methylated nucleic acids **155** and discriminator probes **130** selectively hybridized to converted nucleic acids **150**, is provided by the invention. The methods provided by the invention entail removing converted (i.e., fully unmethylated) subsequences **150** that have hybridized to discriminator probes **130** from methylated (i.e., unhybridized) subsequences **155**, e.g., by any of a variety of methods well known in the art. In preferred embodiments of the methods, converted nucleic acids **150** can be separated from methylated nucleic acids **155** in population **120** e.g., via electrophoresis. In other embodiments, affinity tag **241**, optionally present on discriminator probes **130**, can be advantageously used to separate the duplexes formed by discriminator probes **130** and converted nucleic acids **150** via affinity purification.

[0045] The remaining population of methylated nucleic acids **165** is available for further analyses such as sequencing, e.g., using a high throughput sequencing system. Sequencing enriched population of methylated nucleic acids **165** can provide a map of the methylated subsequences present in nucleic acid **100**.

[0046] Advantageously, the discriminator probes can optionally be reused to probe subsequent populations of nucleic acids in further methylation pattern analyses.

FURTHER DETAILS REGARDING DNA METHYLATION

[0047] The methods and compositions provided by the invention can be beneficially used to determine a "methylation map" of a nucleic acid that comprises both unmethylated and methylated subsequences, e.g., a genomic DNA. The DNA of most organisms is modified by the post-synthetic addition of methyl groups in reactions that are catalyzed by DNA methyltransferase (DNMT) enzymes. In mammals, methylation is almost exclusively found on cytosine C5 of CpG dinucleotides, which are unevenly distributed throughout mammalian genomes in clusters termed CpG islands (CGI). CGIs are typically found in the 5' regions, e.g., in the promoters and/or in first exons, of approximately 50-60% of human genes (Wang, et al. (2004) "An evaluation of new criteria for CpG islands in the human genome as gene markers." *Bioinformatics* **20:** 1170–1177; Larsen, et al. (1992) "CpG islands as gene markers in the human genome." *Genomics* **13:** 1095–1107). However, the methylation profile of less than 0.1% of the human genome has been analyzed in detail (Schumacher, et al. (2006) "Microarray-based DNA methylation profiling: technology and applications." *Nucleic Acids Res* **34:** 528 –542), and thus little is known about methylation

patterns on a genome-wise scale and/or how changes to methylation patterns in a genome impact biological processes such as, e.g., embryonic development, chromatin structure, X chromosome inactivation, genomic imprinting, and chromosome stability. Using the methods and compositions provided herein to generate a high-resolution "methylome" of, e.g., a human genomic DNA, can provide insight into the role that DNA methylation plays in regulating these biological processes.

[0048] DNA methylation profiles in mammals are tissue specific (Khulan, et al. (2006) "Comparative isoschizomer profiling of cytosine methylation: The HELP assay." *Genome Res* **16:** 1046-1055; Kitamura, et al. (2007) "Analysis of tissue-specific differentially methylated regions (TDMs) in humans." *Genomics* **89:** 326–337; Illingworth, et al. (2008) "A novel CpG island set identifies tissue-specific methylation at developmental gene loci." *PLoS Biol* **6:** e22. doi: 10.1371/journal.pbio.0060022). Several tissue-specific differentially methylated regions (tDMRs) have been identified, e.g., in mouse and human genomes, and implicated for their indispensable involvement in mammalian development and tissue differentiation (Ohgane, et al. (2008) "Epigenetics: The DNA Methylation Profile of Tissue-Dependent and Differentially Methylated Regions in Cells" *Placenta* **29:** 29-35; Igarashi, et al. (2008) "Quantitative analysis of human tissue-specific differences in methylation." *Biochem Biophys Res Comm* **376:** 658-664, Song, et al. (2005) "Association of tissue-specific differentially methylated regions (TDMs) with differential gene expression." *Proc Natl Acad Sci U S A* **102:** 3336-3341). However, tissue-specific methylation profiles have not been comprehensively examined in any vertebrate, and the role of DNA methylation in regulating tissue-specific gene expression is poorly understood. Advantageously, a set of discriminator probes can be used to interrogate, e.g., the methylation profiles of genomic DNA derived from two or more different tissues or cell types of one organism, in a high-throughput format. Comparing genome-wide methylation profiles between tissues and cell types can be informative in uncovering the role of DNA methylation in tissue differentiation, e.g., in mammals.

[0049] While genomic methylation patterns are generally stable and heritable in somatic differentiated cells, genome-wide methylation profiles are reprogrammed in mammals during gametogenesis and early embryogenesis, generating cells with a broad developmental potential (Aranyi, et al. (2006) "The constant variation: DNA methylation changes during preimplantation development." *FEBS Lett* **580:** 6521-6526; Marchal, et al.

(2005) "DNA methylation in mouse gametogenesis." *Cytogenet Gen Res* **105:** 316-324). A set of discriminator probes, e.g., provided by the invention, can be used to monitor the temporal patterns of genome methylation, e.g., using methods described herein, during different stages of, e.g., spermatogenesis, oogenesis, embryogenesis, somatic cell nuclear transfer (SCNT), organ development, etc. The methods and compositions of the invention can also aid in identifying sequences targeted for methylation during these events. In addition, methods and compositions of the invention can be used to observe the effects of various environmental factors, e.g., drugs, on methylation dynamics during these events, to provide insight into, e.g., embryogenesis, allele-specific imprinting, X inactivation, chromosome stability, etc.

[0050] The methylation profiles of somatic cells can be altered by, e.g., aging, nutrition, disease, mutational events during embryogenesis, and other factors. Accordingly, alterations in a genome's methylation profile play a role in the pathogenesis of a variety of complex disorders, including, e.g., atherosclerosis, cancer, autoimmune disease, and imprinting disorders, e.g., Prader-Willi and Beckwith/Wiedemann syndromes. The methods and compositions provided by the invention can permit low-cost large-scale analyses in which the methylation profiles of genomic DNA derived from, e.g., normal cells and disease cells, e.g., cancer cells, pancreatic beta cells, or arterial smooth muscle cells, are compared. Cataloging the differences between the methylation profiles of normal cells and disease cells can be useful in identifying informative DNA methylation biomarkers that correlate with or cause disease. Similarly, the methods and compositions can be used to assay genomic DNA derived from the appropriate tissue of a patient in order to determine the methylation pattern at a biomarker in order to, e.g., assess the patient's predisposition for developing the disease, make a diagnosis, predict a patient's prognosis, and/or determine a therapeutic regimen. Relatedly, the methods, e.g., provided by the invention, of using a set of discriminator probes to determine the methylation pattern of a disease-related biomarker can also be used to monitor the efficacy of a therapeutic regimen, e.g., by detecting whether a "normal" methylation pattern at the locus of interest has been restored.

[0051] Further details regarding DNA methylation and its regulation can be found in, e.g., Ng, et al. (2008) "Epigenetic inheritance of cell differentiation status." *Cell Cycle* **7:** 1173-1177; Lees-Murdock, et al. (2008) "DNA methylation reprogramming in the germ line." *Adv Exp Med Biol* **626:** 1-15; Bibikova, et al. (2008) "Unraveling epigenetic

regulation in embryonic stem cells." *Cell Stem Cell* **2:** 123-134; Doerfler and Bohm, eds. DNA Methylation: Basic Mechanisms. USA: Springer, 2006; and others. Details regarding DNA methylation and disease are elaborated in, e.g., Robertson (2005) " DNA methylation and human disease." *Nat Rev Genet* **6:** 597-610; Tost (2009) "DNA methylation: an introduction to the biology and the disease-associated changes of a promising biomarker." *Methods Mol Biol* **507:** 3-20; Manel Esteller, ed. DNA Methylation Epigenetics and Metastasis. USA: Springer, 2005; and others.

FURTHER DETAILS REGARDING HIGH THROUGHPUT SEQUENCING SYSTEMS

[0052] The methylated nucleic acid fragments that are enriched using methods provided by the invention can optionally be sequenced using, e.g., any of a variety of high-throughput DNA sequencing systems (reviewed in, e.g., Chan, et al. (2005) "Advances in Sequencing Technology" (Review) *Mutation Research* **573:** 13-40). See, e.g., Hodges, et al. (2007) "Genome-wide in situ exon capture for selective resequencing." *Nat Genet* **39:** 1522-1527; Olson M (2007) "Enrichment of super-sized resequencing targets from the human genome." *Nat Methods* **4:** 891-892; Porreca, et al. (2007) "Multiplex amplification of large sets of human exons." *Nat Methods* **4:** 931-936.

[0053] One subset of commercial sequencing systems, e.g., those available from Affymetrix and Complete Genomics, Inc., rely on indirect methods of determining a DNA's sequence, e.g., sequencing by hybridization (SBH), in which a sequence of a DNA is assembled based on experimental data obtained from hybridization experiments performed to determine the oligonucleotide content of the DNA chain. See, e.g., Drmanac, et al. (2002) "Sequencing by hybridization (SBH): advantages, achievements, and opportunities." *Adv Biochem Eng Biotechnol* **77:** 75-101. SBH typically employs an array comprising a known arrangement of short oligonucleotides of known sequence, e.g., oligonucleotides representing all possible sequences of a given length. An unknown sequence of, e.g., fluorescently labeled DNA, is fragmented, and the resulting fragments are then hybridized to the oligonucleotide probes in the array. Because the hybridization of a nucleic acid to a short complementary sequence can be sensitive to even single-base mismatches, the hybridization intensity of the labeled nucleic acid fragments to individual probes in the array is computationally assessed to determine the sequences of the fragments. Additional computational approaches are then used to assemble the sequence fragments to determine the entire sequence of the nucleic acid whose fragments were hybridized to the array.

[0054] SoLID, a commercial sequencing system available from Applied Biosystems, is based on "sequencing by ligation" (SBL), in which the mismatch sensitivity of a DNA ligase enzyme is used to determine the underlying sequence of the target nucleic acid molecule. Briefly, one or more sets of encoded adaptors are ligated to the terminus of a target polynucleotide, e.g., a single-stranded DNA of unknown sequence. Encoded adaptors whose protruding strands form perfectly matched duplexes with the complementary protruding strands of the target polynucleotide are ligated, and the identity of the nucleotides in the protruding strands is determined by an oligonucleotide tag carried by the encoded adaptor. Such determination, or "decoding" is carried out by specifically hybridizing a labeled tag complementary to its corresponding tag on the ligated adaptor.

[0055] Other commercial high-throughput sequencing systems, e.g., those available from 454 Life Sciences, Illumina, and Pacific Biosciences, are based on multiplexed direct sequencing methods, e.g., "sequencing by synthesis" (SBS), in which each base position in a single-stranded DNA template is determined individually during the synthesis of a complementary strand. 454 Sequencing, a technology available from 454 Life Sciences, is a massively-parallellized, multiplex pyrosequencing system (Nyren (2007) "The History of Pyrosequencing." *Methods Mol Biol* **373:** 1-14; Ronaghi (2001) "Pyrosequencing sheds light on DNA sequencing." *Genome Res* **11:** 3-11; and Wheeler, et al. (2008) "The complete genome of an individual by massively parallel DNA sequencing." Nature 452: 872-876) that relies on fixing nebulized, adapter-ligated single-stranded DNA fragments to small DNA-capture beads.

[0056] Single molecule real-time sequencing (SMRT) is another massively parallel sequencing technology that can be compatible with the high-throughput resequencing of target nucleic acids isolated isolated from a sample, e.g., by using capture probes synthesized according to any of the methods described previously. Developed and commercialized by Pacific Biosciences, SMRT technology relies on arrays of multiplexed zero-mode waveguides (ZMWs) in which, e.g., thousands of sequencing reactions can take place simultaneously. The ZMW is a structure that creates an illuminated observation volume that is small enough to observe, e.g., the template-dependent synthesis of a single single-stranded DNA molecule by a single DNA polymerase (See, e.g., Levene, et al. (2003) "Zero Mode Waveguides for Single Molecule Analysis at High Concentrations," *Science* **299:** 682-686).

[0057] The methylated sequences that are obtained, e.g., using methods provided by the invention, can be sequenced using any of the systems described above or systems that include bridge amplification technologies, e.g., in which primers bound to a solid phase are used in the extension and amplification of solution phase target nucleic acid acids prior to SBS. (See, e.g., Mercier, et al. (2005) "Solid Phase DNA Amplification: A Brownian Dynamics Study of Crowding Effects." *Biophysical Journal* **89:** 32-42; Bing, et al. (1996) "Bridge Amplification: A Solid Phase PCR System for the Amplification and Detection of Allelic Differences in Single Copy Genes." Proceedings of the Seventh International Symposium on Human Identification, Promega Corporation Madison, WI.) Solexa sequencing, available from Illumina, is one such sequencing system.

FURTHER DETAILS REGARDING MOLECULAR BIOLOGY TECHNIQUES

Subtractive Hybridization

[0058] In the methods provided by the invention, subtractive hybridization is used to separate unmethylated sequences from the methylated sequences in a population of nucleic acids that comprises both methylated and unmethylated sequences, e.g., unconverted and converted fragments of a genomic DNA, respectively. The principle of this approach relies on the hybridization of unmethylated nucleic acid species, e.g., fragments of a genomic DNA, that have been treated with a methylation state conversion reagent, e.g., bisulfite, to a set of discriminator probes that selectively and specifically hybridize to converted nucleotide sequences. The resulting hybridized duplexes are removed, or "subtracted" from the nucleic acid population, leaving methylated sequences for further manipulation and analysis, e.g., sequencing, e.g., using an automated high-throughput sequencing system. These methods can be used to identify, e.g., cell type-specific methylation patterns, tissue-specific methylation patterns, developmental stage-specific methylation patterns, disease-specific methylation patterns, and the like, that are unique to the nucleic acid sample that is being interrogated with the set of discriminator probes.

[0059] The nucleic acid discriminator probes, e.g., produced by methods provided by the invention that are described elsewhere herein, "hybridize" to converted fragments, e.g., unmethylated nucleic acid fragments that have been treated with a methylation state conversion reagent, when they associate, typically in solution. Nucleic acids hybridize due to a variety of well-characterized physico-chemical forces, such as hydrogen bonding, solvent exclusion, base stacking and the like. An extensive guide to the hybridization of

nucleic acids is found in Tijssen (1993) *Laboratory Techniques in Biochemistry and Molecular Biology--Hybridization with Nucleic Acid Probes* part I chapter 2, "Overview of principles of hybridization and the strategy of nucleic acid probe assays," (Elsevier, New York), as well as in Current Protocols in Molecular Biology, Ausubel et al., eds., Current Protocols, a joint venture between Greene Publishing Associates, Inc. and John Wiley & Sons, Inc., (supplemented through 2004) ("Ausubel"); Hames and Higgins (1995) Gene Probes 1 IRL Press at Oxford University Press, Oxford, England, (Hames and Higgins 1) and Hames and Higgins (1995) Gene Probes 2 IRL Press at Oxford University Press, Oxford, England (Hames and Higgins 2).

[0060] In general, the stringency of the conditions under which, e.g., discriminator probes and converted fragments, are hybridized, e.g., in the methods described herein, are experimentally determined. An extensive guide to the hybridization of nucleic acids is found in Tijssen (1993), *supra*, and in Hames and Higgins, 1 and 2. In certain embodiments of the methods, hexadecyltrimethylammonium bromide can be added to a hybridization mix that includes the discriminator probes and converted nucleic acid fragments to increase the specificity and kinetics of hybridization.

[0061] Typically, the results of subtractive hybridization are validated using additional techniques that are well known in the art, e.g., northern blot, *in situ* hybridization, RT-PCR, and the like. These techniques are described in detail in, e.g., e.g., Sambrook et al., Molecular Cloning - A Laboratory Manual (3rd Ed.), Vol. 1-3, Cold Spring Harbor Laboratory, Cold Spring Harbor, New York, 2000 ("Sambrook"); and Current Protocols in Molecular Biology, F.M. Ausubel et al., eds., Current Protocols, a joint venture between Greene Publishing Associates, Inc. and John Wiley & Sons, Inc., (supplemented through 2007) ("Ausubel").

[0062] Further details regarding subtractive hybridization are elaborated in, e.g., Aasheim, et al. (1996) "Subtractive hybridization for the isolation of differentially expressed genes using magnetic beads." *Meth Mol Biol* **69:** 115-128; Wink. "The Investigation of Transcriptional Activity." *An Introduction to Molecular Biotechnology.* Ed. Michael Wink. Germany: Wiley-VCH, 2006. 334-340; Blumberg, et al. "Subtractive Hybridization and Construction of cDNA Libraries." *Methods in Molecular Biology, Vol. 97.* Ed. Sharpe and Mason. Totowa, NJ: Humana Press, Inc, 1999. 119-129; Røsok, et al. "Discovery of differentially expressed genes: technical considerations." *Methods in*

Molecular Biology, Vol. 360. Ed. Sharpe and Mason. Totowa, NJ: Humana Press, Inc, 2007. 115-129; Darwin (2005) "Genome-wide screens to identify genes of human pathogenic *Yersinia* species that are expressed during host infection." *Curr Issues Mol Biol* **7:** 135-49; and others. In addition, subtractive hybridization kits are commercially available, including PCR-Select™ cDNA Subtraction Kit (Clontech).

Preparing genomic DNA

[0063] As described above, both the set of discriminator probes and the population of nucleic acids to which the probes are hybridized, e.g., for the purpose of separating methylated sequences in the population from unmethylated sequences, can be derived from a genomic DNA. Genomic DNA can be prepared from any source by three steps: cell lysis, deproteinization and recovery of DNA. These steps are adapted to the demands of the application, the requested yield, purity and molecular weight of the DNA, and the amount and history of the source. Further details regarding the isolation of genomic DNA can be found in Berger and Kimmel, Guide to Molecular Cloning Techniques, Methods in Enzymology volume 152 Academic Press, Inc., San Diego, CA (Berger); Sambrook et al., Molecular Cloning - A Laboratory Manual (3rd Ed.), Vol. 1-3, Cold Spring Harbor Laboratory, Cold Spring Harbor, New York, 2008 ("Sambrook"); Current Protocols in Molecular Biology, F.M. Ausubel et al., eds., Current Protocols, a joint venture between Greene Publishing Associates, Inc. and John Wiley & Sons, Inc ("Ausubel"); Kaufman et al. (2003) Handbook of Molecular and Cellular Methods in Biology and Medicine Second Edition Ceske (ed) CRC Press (Kaufman); and The Nucleic Acid Protocols Handbook Ralph Rapley (ed) (2000) Cold Spring Harbor, Humana Press Inc (Rapley). In addition, many kits are commercially available for the purification of genomic DNA from cells, including Wizard™ Genomic DNA Purification Kit, available from Promega; Aqua Pure™ Genomic DNA Isolation Kit, available from BioRad; Easy-DNA™ Kit, available from Invitrogen; and DnEasy™ Tissue Kit, which is available from Qiagen.

Generating Nucleic Acid Fragments

[0064] In the methods described herein, nucleic acid fragments comprising both methylated and unmethylated sequences are generated, e.g., from a genomic DNA, i.e., in preparation for treatment with a methylation state conversion reagent and hybridization to a set of discriminator probes, i.e., to facilitate the removal of unmethylated sequences from

the methylated sequences. There exist a plethora of ways of producing such nucleic acid fragments. These include, but are not limited to, mechanical methods, such as sonication, mechanical shearing, nebulization, hydroshearing, and the like; enzymatic methods, such as exonuclease digestion, endonuclease digestion, and the like; chemical cleavage, and electrochemical cleavage. These methods are further explained in Sambrook and Ausubel.

Nucleic acid tags

[0065] The invention provides compositions that include a set of tagged discriminator probes that selectively hybridize to unmethylated sequences in a population of nucleic acids. The tags can permit the detection of nucleic acid duplexes comprising a discriminator probe and an unmethylated sequence, e.g., in a population of nucleic acids that comprises a subpopulation of methylated sequences and a subpopulation of unmethylated sequences, e.g., following hybridization of probes to a population of nucleic acid fragments derived from, e.g., a genomic DNA. In addition, the tags permit the aforementioned nucleic acid duplexes to be separated, e.g., via affinity purification or the like, from the subpopulation of unhybridized nucleic acid fragments, e.g., fragments that contain methylated sequences. Nucleic acid tags, e.g., such as those optionally present on the discriminator probes, can comprise any of a plethora of ligands, such as high-affinity DNA-binding proteins; modified nucleotides, such as methylated, biotinylated, or fluorinated nucleotides; and nucleotide analogs, such as dye-labeled nucleotides, non-hydrolysable nucleotides, or nucleotides comprising heavy atoms. For example, tags can optionally comprise one or more fluorescent label, blocking group, phosphorylated nucleotide, thiol linker, phosphorothioated nucleotide, amine-reactive nucleotide, uracil, and/or the like. Such reagents are widely available from a variety of vendors, including Perkin Elmer, Jena Bioscience and Sigma-Aldrich.

[0066] Nucleic acid tags can also include oligonucleotides that comprise specific sequences, such as restriction sites, cis regulatory sites, nucleotide hybridization sites, protein binding sites, sequences capable of forming hairpin secondary structures, DNA promoters, sample or library identification sequences, and the like. Such sequences can be of advantageous use in, e.g., sequencing tagged methylated nucleic acids derived from a population of nucleic acids which comprises both methylated and unmethylated subsequences. Linkers that are attached to methylated sequences in preparation for

sequencing, e.g., in a high-throughput sequencing system, can also beneficially include any one or more of the sequences listed above. Oligonucleotide tags can be custom synthesized by commercial suppliers such as Operon (Huntsville, AL), IDT (Coralville, IA) and Bioneer (Alameda, CA). Any of a number of methods that are well known in the art can be used to join tags to nucleic acids of interest, include chemical linkage, ligation, and extension of a primer comprising a tag by a polymerase or reverse transcriptase. Further details regarding nucleic acid tags and the methods by which they are attached to nucleic acids of interest are elaborated in Sambrook and Ausubel.

Amplifying and copying nucleic acids

[0067] Methods for producing a set of discriminator probes that selectively hybridize to unmethylated sequences that have been treated with a conversion reagent include an amplification step and a copying step (see **Figure 2** and corresponding description). A variety of nucleic acid amplification and/or copying methods are known in the art and can be implemented to, e.g., amplify nucleic acid fragments to produce unmethylated fragments during probe production or to, e.g., copy converted fragments to generate a set discriminator probes. The most widely used *in vitro* technique among these methods is polymerase chain reaction (PCR), which requires the addition of nucleotides, oligonucleotide primers, buffer, and an appropriate polymerase to the amplification reaction mix. Additional methods that can be used to amplify, or copy, nucleic acids include strand displacement amplification (SDA), rolling-circle amplification (RCA) and multiple-displacement amplification (MDA). Each of these techniques is further described in Sambrook et al., Molecular Cloning - A Laboratory Manual (3rd Ed.), Vol. 1-3, Cold Spring Harbor Laboratory, Cold Spring Harbor, New York, 2000 ("Sambrook"); and Current Protocols in Molecular Biology, F.M. Ausubel et al., eds., Current Protocols, a joint venture between Greene Publishing Associates, Inc. and John Wiley & Sons, Inc., (supplemented through 2007) ("Ausubel"), and DNA Amplification: Current Technologies and Applications, V. V. Demidov et al., eds., (1st Ed.), Taylor and Francis, 2004.

KITS

[0068] Kits are also a feature of the invention. The present invention provides kits that include useful reagents, e.g., tagged DNA primers, affinity columns, and/or one or more enzyme and/or reagent that are used in the methods, e.g., a DNA polymerase, bisulfite, etc. Such reagents are most preferably packaged in a fashion to enable their use. The kits of the

invention optionally include additional reagents, such as a control target nucleic acids, buffer solutions and/or salt solutions, including, e.g., divalent metal ions, i.e., Mg^{++}, Mn^{++} and/or Fe^{++}, nucleic acid adapter tags, e.g., to prepare methylated nucleic acid fragments for sequencing, e.g., using a currently available or future automated high-throughput sequencing system. Such kits also typically include a container to hold the kit components, and instructions for use of the compositions, e.g., to practice the methods.

SYSTEMS

[0069] The methods and compositions provided by the invention can advantageously be integrated with systems that can, e.g., automate and/or multiplex the steps of the methods described herein, e.g., methods for separating methylated subsequences from unmethylated subsequences in a population of nucleic acids. Systems of the invention can include one or more modules, e.g., that automate a method herein, e.g., for high-throughput sequencing applications. Such systems can include fluid-handling elements and controllers that move reaction components into contacts with one another, signal detectors, and system software/instructions.

[0070] Systems of the invention can optionally include modules that provide for detection or tracking of products, e.g., methylated sequences. Additionally or alternatively, the systems can detect the nucleotide sequence of such methylated nucleic acids, e.g., produced during a sequencing reaction. Detectors can include spectrophotometers, epifluorescent detectors, CCD arrays, CMOS arrays, microscopes, cameras, or the like. Optical labeling is particularly useful because of the sensitivity and ease of detection of these labels, as well as their relative handling safety, and the ease of integration with available detection systems (e.g., using microscopes, cameras, photomultipliers, CCD arrays, CMOS arrays and/or combinations thereof). High-throughput analysis systems using optical labels include DNA sequencers, array readout systems, cell analysis and sorting systems, and the like. For a brief overview of fluorescent products and technologies see, e.g., Sullivan (ed) (2007) Fluorescent Proteins, Volume 85, Second Edition (Methods in Cell Biology) (Methods in Cell Biology) ISBN-10: 0123725585; Hof et al. (eds) (2005) Fluorescence Spectroscopy in Biology: Advanced Methods and their Applications to Membranes, Proteins, DNA, and Cells (Springer Series on Fluorescence) ISBN-10: 354022338X; Haughland (2005) Handbook of Fluorescent Probes and Research Products, 10th Edition (Invitrogen, Inc./ Molecular Probes); BioProbes Handbook, (2002) from

Molecular Probes, Inc.; and Valeur (2001) Molecular Fluorescence: Principles and Applications Wiley ISBN-10: 352729919X. System software, e.g., instructions running on a computer can be used to track and inventory reactants or products, and/or for controlling robotics/ fluid handlers to achieve transfer between system stations/modules. The overall system can optionally be integrated into a single apparatus, or can consist of multiple apparatus with overall system software/instructions providing an operable linkage between modules.

EXAMPLES

[0071] The following examples are offered to illustrate, but not to limit the claimed invention.

Preparing Discriminator Probes from a Genomic DNA

[0072] To prepare discriminator probes, 10μg of genomic DNA is sonicated to produce nucleic acid fragments of an average size of ~2kb. The sonicated DNA is then purified using a spin column included with the Qiagen QIAquick PCR Purification Kit and eluted in 50μl hot (>65°C) Buffer EB.

[0073] The following reagents are added to the DNA fragments:

10μl 10X T4 DNA Ligase Buffer
4μl T4 DNA Polymerase (NEB 3U/μl)
5μl PNK (NEB or Fermentas)
2μl 10mM dNTPs
31μl Water

The mixture is incubated at 20°C for 15 minutes and the DNA is purified using a Qiagen spin column as described above. The phosphorylated fragments are eluted in 50μl hot (>65°C) EB.

[0074] The following reagents are added to the DNA recovered from the previous step:

6μl 10X Thermopol Buffer (NEB)
1μl 100mM dATP
1μl Taq Polymerase
4μl water

This reaction mix is incubated at 72°C for 1 hour and purified using a Qiagen spin column, as described above.

[0075] Solexa adapters are ligated to the ends of the fragments produced above by adding the following reagents to the fragments:

6μl 10X T4 DNA Ligase Buffer
1μl SolA-TA (2.5μM)
2μl T4 DNA Ligase (High Concentration)
3μl Water

This reaction mix is purified using a Qiagen spin column, as described above.

[0076] The tagged fragments, e.g., fragments to which the Solexa adapters have been ligated, are then amplified in the following manner:

To 20μl tagged DNA fragments, add:
5μl 10X Taq polymerase buffer
5μl solexa A primer (2pM)
1μl dNTP
18μl water
1μl Taq

The mixture is heated to 95°C for 5 min, cooled to 60°C to permit primer annealing, and heated to 72°C to permit the polymerase to extend the annealed primers. This cycle is repeated 3 times, and the amplified DNA is then purified using a Qiagen spin column, as described above.

[0077] The amplified fragments are treated with bisulfite using a kit available from Qiagen according to manufacturer's instructions. The converted fragments are amplified again in the following manner:

To 50μl converted DNA, add:
10μl 10X Phi29 Buffer
10μl 10X BSA
2μl Biotin-solexa A-converted primer (2pM)
17μl water

The reaction mixture is heated to 95°C for 5 min to denature the DNA, then cooled to at 65°C for 5 min to permit primer annealing. The mixture is then cooled to room temperature. 1μl Phi29 polymerase is added to the reaction, and the reaction mix is incubated at 30°C for 10 minutes. The amplification reaction is then purified using a Qiagen spin column, as described above.

Preparing a Population of Sample Nucleic Acids

[0078] 10μg of genomic DNA are fragmented to produce fragments of an average length of ~50bp. The fragments are treated with bisulfite, as described above, and following bisulfite conversion, the fragments are purified using the QIAquick Nucleotide Removal Kit according to manufacturer's instructions.

Hybridizing Discriminator Probes to Sample Nucleic Acids

[0079] The following reagents are mixed together and incubated at 95°C for 10 minutes:

Half of the probe DNA (~5μg)
3μg sample DNA
10μl NEB Buffer 2
add 1mM CTAB
water up to 100μl

The mixture is then slowly cooled to 65°C and incubated at this temperature overnight.

[0080] Beads are prepared in the following manner: 10-20 ng (equivalent to 10-20μl) Dynal Beads are washed twice in 2x B&W buffer and resuspended in 100μl 2x B&W. 100μl of biotin DNA in water is added to the beads and allowed to bind for 15 minutes. The beads are pelletted, and the supernatant is transferred to a 1.5ml tube. The beads are washed twice in 50μl 1X NEB buffer 2. The supernatants are from each wash are saved, and the beads are discarded.

[0081] The following reagents are mixed together, incubated at 95°C for 10 minutes, slowly cooled to 65°C, and incubated overnight:

~5μg discriminator probe DNA
sample DNA
20μl NEB Buffer 2
1mM CTAB
water up to 200μl

[0082] After preparing beads as described above,

Add 200μl of biotin DNA in water
Bind for 15 min
Pellet beads
Transfer solution to a 1.5ml tube
Wash beads 2x in 50μl 1X NEB buffer 2.

The supernatants from the washes are saved and the beads discarded. The unhybridized DNA fragments are purified using a QIAquick Nucleotide Removal Kit. Solexa adapters are added to the ends of the fragments with RNA ligase and sequenced.

[0083] While the foregoing invention has been described in some detail for purposes of clarity and understanding, it will be clear to one skilled in the art from a reading of this disclosure that various changes in form and detail can be made without departing from the true scope of the invention. For example, all the techniques and apparatus described above can be used in various combinations. All publications, patents, patent applications, and/or other documents cited in this application are incorporated by reference in their entirety for all purposes to the same extent as if each individual publication, patent, patent application, and/or other document were individually indicated to be incorporated by reference for all purposes.

CLAIMS

WHAT IS CLAIMED IS:

1. A method of distinguishing unmethylated subsequences from methylated subsequences in a nucleic acid sample, the method comprising:

a) providing the nucleic acid sample;

b) fragmenting the nucleic acid sample to produce fragments;

c) treating the fragments with a methylation state conversion reagent to produce treated nucleic acids that comprise a subset of converted nucleic acids and a subset of unconverted nucleic acids, wherein the converted nucleic acids correspond to the unmethylated subsequences in the nucleic acid sample and wherein the unconverted nucleic acids correspond to methylated subsequences in the nucleic acid sample; and,

d) hybridizing the treated nucleic acids to a set of discriminator probes that selectively hybridize to the converted nucleic acids to produce hybridized nucleic acids, thereby distinguishing unmethylated subsequences from methylated subsequences in a nucleic acid sample.

2. The method of claim **1**, wherein the nucleic acid sample and the discriminator probes are both derived from a first source of nucleic acids.

3. The method of claim **2**, wherein the first source of nucleic acids is derived from a genomic DNA.

4. The method of claim **1**, wherein fragmenting the nucleic acid sample comprises one or more methods selected from: enzymatic digestion, sonication, mechanical shearing, electrochemical cleavage, and nebulization.

5. The method of claim **1**, wherein the methylation state conversion reagent comprises sodium bisulfite.

6. The method of claim **1**, wherein the method further comprises separating the hybridized nucleic acids from the unconverted nucleic acids.

7. The method of claim **6**, wherein separating the hybridized nucleic acids from the unconverted nucleic acids comprises:

a) hybridizing the converted nucleic acids to the discriminator probes, wherein the discriminator probes comprise tags, to produce tagged hybridized nucleic acids; and,

b) removing the tagged hybridized nucleic acids from the unconverted nucleic acids via affinity purification.

8. The method of claim **7**, wherein the tags comprise one or more moieties selected from: a ligand, a fluorescent label, a blocking group, a phosphorylated nucleotide, a biotinylated nucleotide, a methylated nucleotide, a uracil, a sequence capable of forming hairpin secondary structure, an oligonucleotide hybridization site, a restriction site, a DNA promoter, a protein binding sequence, a sample or library identification sequence, a thiol linker, a phosphorothioated nucleotide, an amine reactive nucleotide, and a cis regulatory sequence.

9. The method of claim **6**, wherein separating the hybridized nucleic acids from the unconverted nucleic acids fragments comprises electrophoresis.

10. The method of claim **6**, wherein the method further comprises sequencing the unconverted nucleic acids and comparing sequences of the unconverted nucleic acids to sequences of the nucleic acid sample to identify the methylated subsequences in the nucleic acid sample.

11. The method of claim **10**, wherein the unhybridized nucleic acids are sequenced by a high-throughput sequencing system.

12. A composition comprising a set of nucleic acids that comprises methylated subsequences of a genomic DNA, which set has been produced by:

a) providing a genomic DNA;

b) fragmenting the genomic DNA to produce genomic fragments;

c) treating the genomic fragments with a methylation state conversion reagent to produce treated nucleic acids that comprise a subset of converted nucleic acids and a subset

of unconverted nucleic acids, wherein the converted nucleic acids correspond to the unmethylated subsequences in the nucleic acid sample and wherein the unconverted nucleic acids correspond to the methylated subsequences in the nucleic acid sample;

d) hybridizing the treated nucleic acids to a set of discriminator probes that selectively hybridize to the converted nucleic acids to produce hybridized nucleic acids; and,

e) separating the hybridized nucleic acids from the unconverted nucleic acids, thereby separating the unmethylated subsequences from the methylated subsequences in the genomic DNA.

13. A composition, comprising:

a) a set of discriminator probes that selectively hybridize to converted nucleic acids in a population of nucleic acids that has been treated with a methylation state conversion reagent; and,

b) a population of nucleic acids that comprises a subset of methylated nucleic acids and a subset of unmethylated nucleic acids, which population has been treated with the conversion reagent to produce a subset of converted nucleic acids and a subset of unconverted nucleic acids, wherein the converted nucleic acids correspond to the unmethylated nucleic acids, and wherein the unconverted nucleic acids correspond to the methylated nucleic acids,

wherein the converted nucleic acids are hybridized to the set of discriminator probes.

14. The composition of claim **13**, wherein the discriminator probes and the population of nucleic acids are both derived from a first source of nucleic acids.

15. The composition of claim **13**, wherein the first source of nucleic acids is derived from a genomic DNA.

16. The composition of claim **13**, wherein the methylation state conversion reagent is sodium bisulfite.

17. A composition comprising a set of probes capable of distinguishing between unmethylated nucleic acid sequences and methylated nucleic acid sequences.

18. A method of producing a set of discriminator probes that selectively hybridize to converted sequences in a nucleic acid sample that has been treated with a methylation state conversion reagent, the method comprising:

a) providing at least one nucleic acid that corresponds to a sequence present in the nucleic acid sample, which sequence comprises both methylated subsequences and unmethylated subsequences;

b) fragmenting the at least one nucleic acid to produce fragments;

c) amplifying the fragments to produce a population of unmethylated nucleic acids;

d) treating the population of unmethylated nucleic acids with the methylation state conversion reagent to produce converted nucleic acids; and,

e) copying the converted nucleic acids to produce the set of discriminator probes, which probes selectively hybridize to converted sequences in the nucleic acid sample that has been treated with the methylation state conversion reagent.

19. The method of claim **18**, wherein the at least one nucleic acid and the nucleic acid sample are both derived from a first source of nucleic acids.

20. The method of claim **19**, wherein the first source of nucleic acids is derived from a genomic DNA.

21. The method of claim **18**, wherein fragmenting the at least one nucleic acid comprises one or more methods selected from: enzymatic digestion, sonication, mechanical shearing, electrochemical cleavage, and nebulization.

22. The method of claim **18**, wherein amplifying the fragments comprises one or more methods selected from: PCR and primer extension.

23. The method of claim **18**, wherein the methylation state conversion reagent comprises sodium bisulfite.

24. The method of claim **18**, wherein copying the converted nucleic acids comprises annealing DNA primers comprising tags to 3' ends of the converted nucleic acids and extending the tagged primers with a polymerase.

25. The method of claim **24**, wherein the tags comprise one or more moieties selected from: a ligand, a fluorescent label, a blocking group, a phosphorylated nucleotide, a biotinylated nucleotide, a methylated nucleotide, a uracil, a sequence capable of forming hairpin secondary structure, an oligonucleotide hybridization site, a restriction site, a DNA promoter, a protein binding sequence, a sample or library identification sequence, a thiol linker, a phosphorothioated nucleotide, an amine reactive nucleotide, and a cis regulatory sequence.

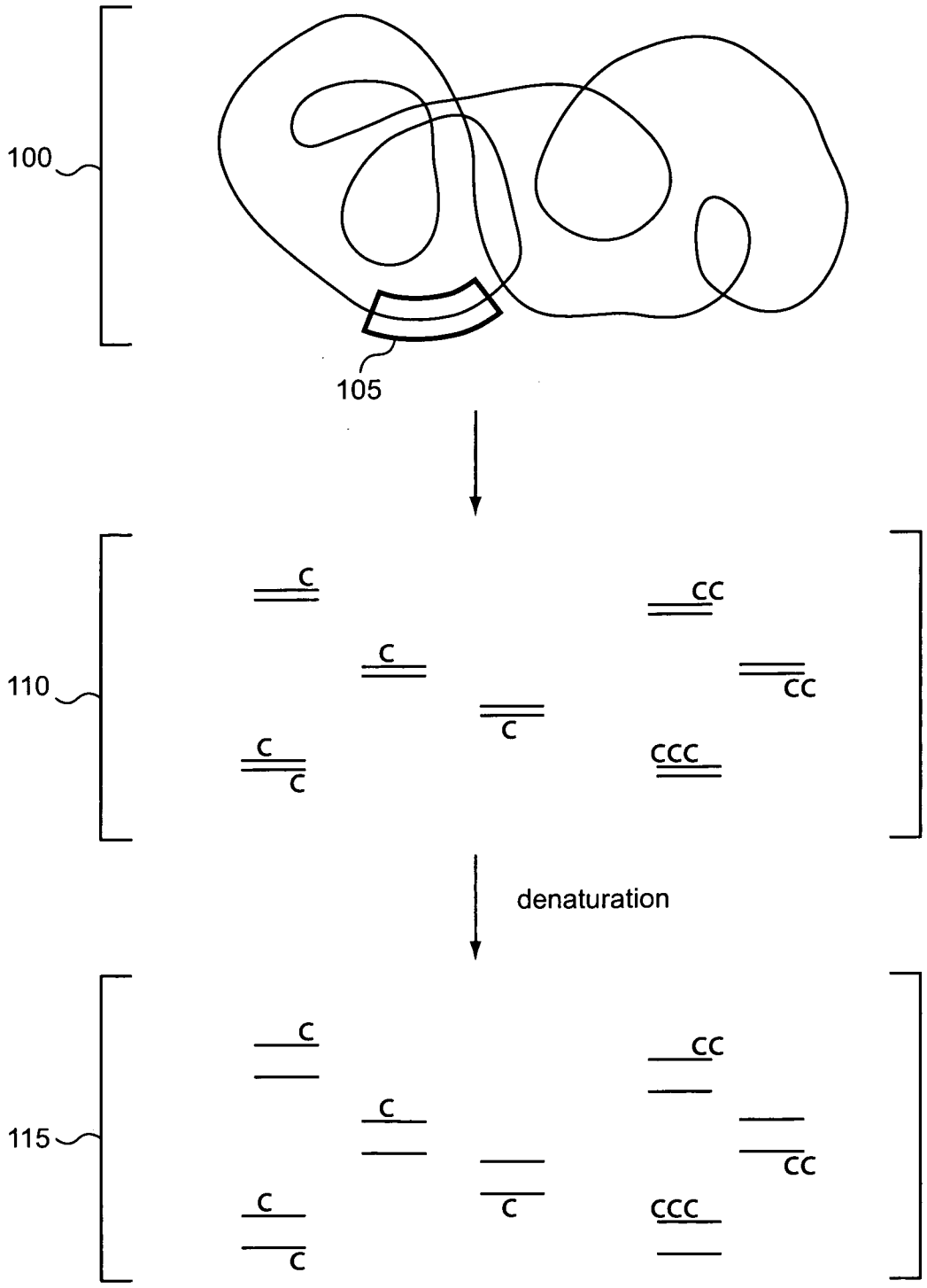


Fig. 1

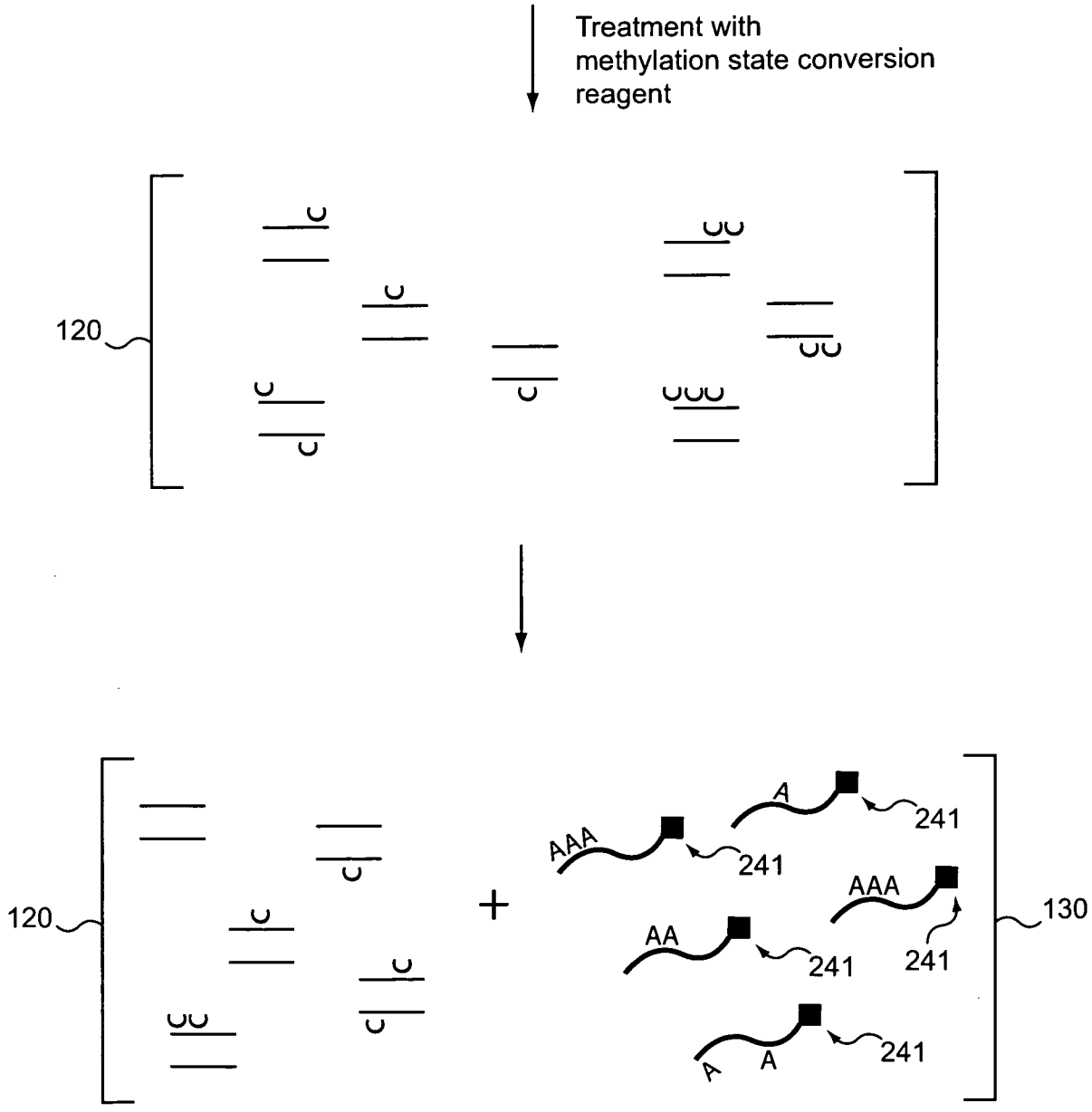

Treatment with
methylation state conversion
reagent
120
120
130
241
241
241
241
241
AAA
AAA
AA
A
A
A

Fig. 1 (cont'd)

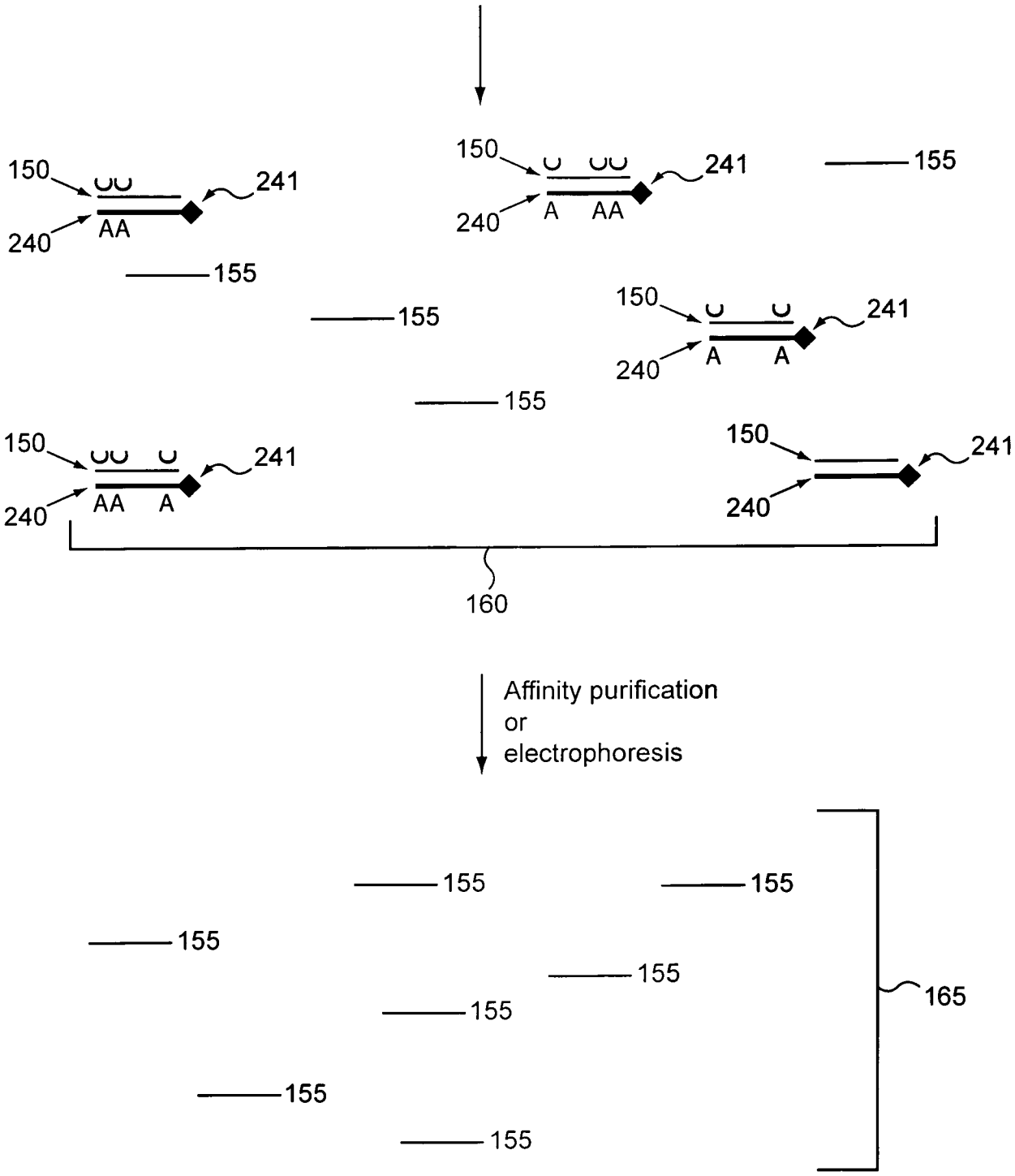

150
241
240
AA
155
150
241
240
A AA
155
155
150
241
240
A A
155
150
241
240
AA A
150
241
240
160
Affinity purification
or
electrophoresis
155
155
155
155
155
155
165

Fig. 1 (cont'd)

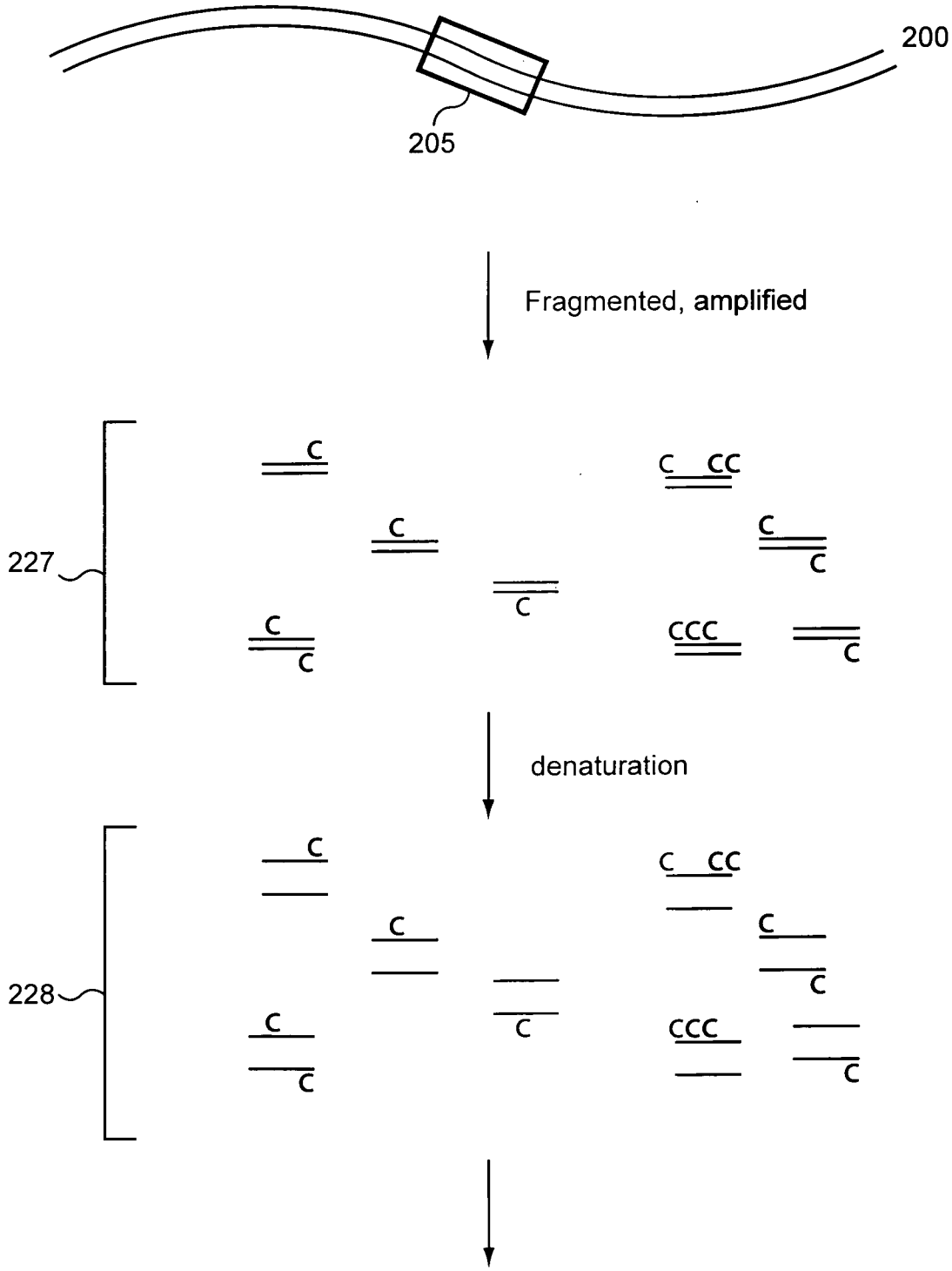


Fig. 2

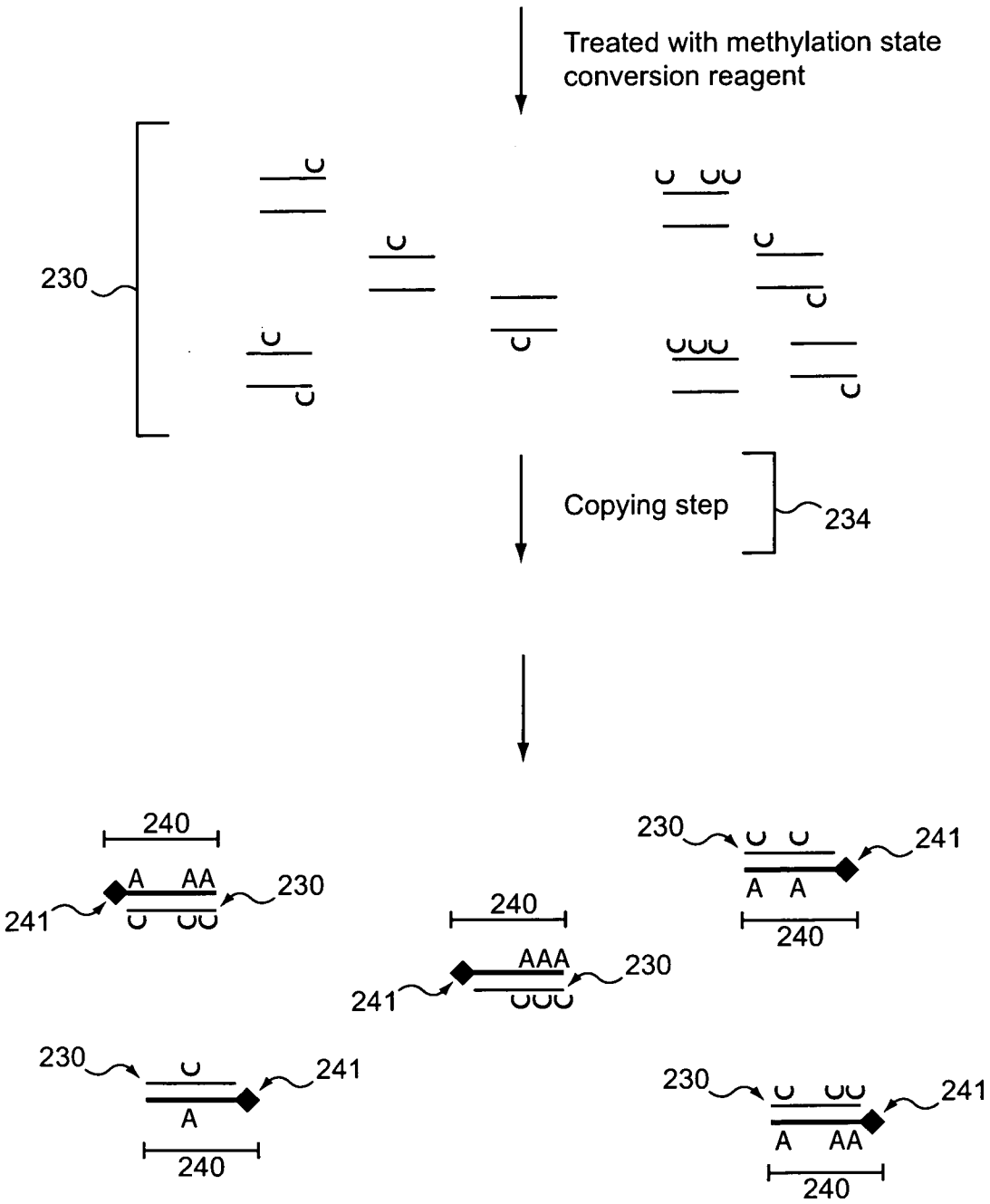

Treated with methylation state conversion reagent
230
Copying step
234
240
241
230
A AA
240
241
230
AAA
230
241
A A
240
230
241
A
240
230
241
A AA
240

Fig. 2 (cont'd)

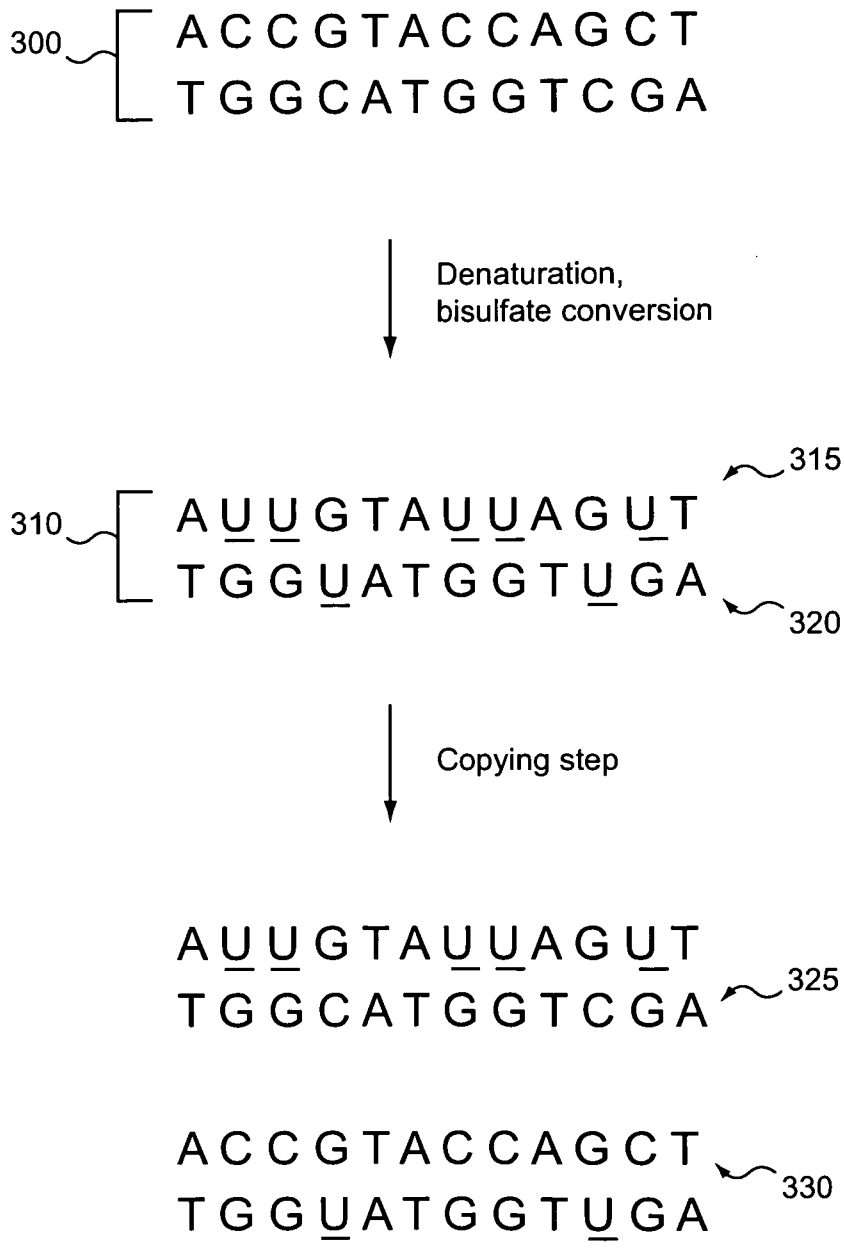

300
ACCGTACCAGCT
TGGCATGGTCGA
Denaturation,
bisulfate conversion
315
310
AUUGTAUUAGUT
TGGUATGGTUGA
320
Copying step
AUUGTAUUAGUT
TGGCATGGTCGA
325
ACCGTACCAGCT
TGGUATGGTUGA
330

Fig. 3