spatially separated reactions. Furthermore, the patent and technical literature includes many examples of microfluidic systems which can interface directly with microwell plates for fluid handling.

Desirable Properties

[0231] The polymerases of the invention can include any of a variety of modified properties towards natural or nucleotide analogs or analogs, depending on the application, including increased branching fractions, delayed translocation, increased nucleotide or nucleotide analog residence time, greater processivity, etc. For example, branching rates can be directly monitored in high-throughput SMS reactions using known templates. Branching or translocation rates can be screened for or against in selecting a polymerase of the invention, e.g., by screening enzymes based on kinetic or product formation properties. Nucleotide analog residence time is readily determined by observing signal pulse widths as detected by a signal detector of a sequencing system.

[0232] For example, improvements in a dissociation rate (or improved processivity) of 30% or more, e.g., about 50%, 75%, or even 100% or more can be screened for in identifying polymerases that display decreased translocation rates. Similarly, detecting mutant polymerases that exhibit branch rates of more than 30%, e.g., 40% or more, preferably 50% or more, or even 75% or more is a feature of the invention.

AFFINITY TAGS AND OTHER OPTIONAL POLYMERASE FEATURES

[0233] The recombinant DNA polymerases optionally include additional features exogenous or heterologous to the polymerases. For example, the recombinant polymerases optionally include one or more exogenous affinity tags, e.g., purification or substrate binding tags, such as a GST tag, an HA tag sequence, a plurality of GST tags, a plurality of HA tag sequences, a SNAP-tag, a c-myc tag, a c-myc fusion, or the like. These and other features useful in the context of binding a polymerase to a surface are optionally included, e.g., to orient and/or protect the polymerase active site when the polymerase is bound to a surface. Other useful features include recombinant dimer domains of the enzyme, and, e.g., large extraneous polypeptide domains coupled to the polymerase distal to the active site. For example, for Φ29, the active site is in the C terminal region of the protein, and added surface binding elements

PACIFIC BIOSCIENCES OF CALIFORNIA, INC. EXHIBIT - 1005 - Part 2
Pg.384

(extra domains, GST tags, etc.) are typically located in the N-terminal region to avoid interfering with the active site when the polymerase is coupled to a surface.

[0234] In general, surface binding elements and purification tags that can be added to the polymerase (recombinantly or, e.g., chemically) include, e.g., biotin, avidin, GST sequences, modified GST sequences, e.g., that are less likely to form dimers, BiTag sequences, S tags, SNAP-tags, enterokinase sites, thrombin sites, antibodies or antibody domains, antibody fragments, antigens, receptors, receptor domains, receptor fragments, ligands, dyes, acceptors, quenchers, or combinations thereof.

[0235] Multiple surface binding domains can be added to orient the polypeptide relative to a surface and/or to increase binding of the polymerase to the surface. By binding a surface at two or more sites, through two or more separate tags, the polymerase is held in a relatively fixed orientation with respect to the surface. Additional details on fixing a polymerase to a surface, attaching tags, and the like are found in WO 2007/075987 ACTIVE SURFACE COUPLED POLYMERASES by Hanzel et al., and WO 2007/076057 PROTEIN ENGINEERING STRATEGIES TO OPTIMIZE ACTIVITY OF SURFACE ATTACHED PROTEINS by Hanzel et al.

MAKING AND ISOLATING RECOMBINANT POLYMERASES

[0236] Generally, nucleic acids encoding a polymerase of the invention can be made by cloning, recombination, in vitro synthesis, in vitro amplification and/or other available methods. A variety of recombinant methods can be used for expressing an expression vector that encodes a polymerase of the invention, e.g., a modified polymerase with an increased branching rate, delayed translocation or increased nucleotide analog residence time. Recombinant methods for making nucleic acids, expression and isolation of expressed products are well known and described in the art. For example, when modifying the active site to increase branching properties, features are selected (e.g., by modeling, though random approaches can also be used) that hinder steric access of the nucleotide analog to the active site, and/or that interfere with charge-charge or hydrophobic interactions between a given nucleotide analog and the polymerase target. Methods for making and selecting mutations in the active site of polymerases, including for modifying steric features in or near the active site to permit

improved access by nucleotide analogs are found, e.g., in WO 2007/076057 POLYMERASES FOR NUCLEOTIDE ANALOG INCORPORATION by Hanzel et al., and PCT/US2007/022459 POLYMERASE ENZYMES AND REAGENTS FOR ENHANCED NUCLEIC ACID SEQUENCING by Rank et al.

[0237] Additional useful references for mutation, recombinant and in vitro nucleic acid manipulation methods (including cloning, expression, PCR, and the like) include Berger and Kimmel, Guide to Molecular Cloning Techniques, Methods in Enzymology volume 152 Academic Press, Inc., San Diego, CA (Berger); 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); Chen et al. (ed) PCR Cloning Protocols, Second Edition (Methods in Molecular Biology, volume 192) Humana Press; and in Viljoen et al. (2005) Molecular Diagnostic PCR Handbook Springer, ISBN 1402034032.

[0238] In addition, a plethora of kits are commercially available for the purification of plasmids or other relevant nucleic acids from cells, (*see*, e.g., EasyPrep™, FlexiPrep™, both from Pharmacia Biotech; StrataClean™, from Stratagene; and, QIAprep™ from Qiagen). Any isolated and/or purified nucleic acid can be further manipulated to produce other nucleic acids, used to transfect cells, incorporated into related vectors to infect organisms for expression, and/or the like. Typical cloning vectors contain transcription and translation terminators, transcription and translation initiation sequences, and promoters useful for regulation of the expression of the particular target nucleic acid. The vectors optionally comprise generic expression cassettes containing at least one independent terminator sequence, sequences permitting replication of the cassette in eukaryotes, or prokaryotes, or both, (e.g., shuttle vectors) and selection markers for both prokaryotic and eukaryotic systems. Vectors are suitable for replication and integration in prokaryotes, eukaryotes, or both.

[0239] Other useful references, e.g. for cell isolation and culture (e.g., for subsequent nucleic acid isolation) include Freshney (1994) Culture of Animal Cells, a Manual of Basic Technique, third edition, Wiley- Liss, New York and the references cited therein; Payne *et al.* (1992) Plant Cell and Tissue Culture in Liquid Systems John Wiley & Sons, Inc. New York,

NY; Gamborg and Phillips (eds) (1995) Plant Cell, Tissue and Organ Culture; Fundamental Methods Springer Lab Manual, Springer-Verlag (Berlin Heidelberg New York) and Atlas and Parks (eds) The Handbook of Microbiological Media (1993) CRC Press, Boca Raton, FL.

[0240] A variety of protein isolation and detection methods are known and can be used to isolate polymerases, e.g., from recombinant cultures of cells expressing the recombinant polymerases of the invention. A variety of protein isolation and detection methods are well known in the art, including, e.g., those set forth in R. Scopes, Protein Purification, Springer-Verlag, N.Y. (1982); Deutscher, Methods in Enzymology Vol. 182: Guide to Protein Purification, Academic Press, Inc. N.Y. (1990); Sandana (1997) Bioseparation of Proteins, Academic Press, Inc.; Bollag et al. (1996) Protein Methods, 2nd Edition Wiley-Liss, NY; Walker (1996) The Protein Protocols Handbook Humana Press, NJ, Harris and Angal (1990) Protein Purification Applications: A Practical Approach IRL Press at Oxford, Oxford, England; Harris and Angal Protein Purification Methods: A Practical Approach IRL Press at Oxford, Oxford, England; Scopes (1993) Protein Purification: Principles and Practice 3rd Edition Springer Verlag, NY; Janson and Ryden (1998) Protein Purification: Principles, High Resolution Methods and Applications, Second Edition Wiley-VCH, NY; and Walker (1998) Protein Protocols on CD-ROM Humana Press, NJ; and the references cited therein. Additional details regarding protein purification and detection methods can be found in Satinder Ahuja ed., Handbook of Bioseparations, Academic Press (2000).

NUCLEIC ACID AND POLYPEPTIDE SEQUENCE AND VARIANTS

[0241] As described herein, the invention provides polynucleotide sequences encoding, e.g., a polymerase as described herein. Examples of polymerase sequences that include features are found herein, e.g., increased branching fractions as in Table A. However, one of skill in the art will immediately appreciate that the invention is not limited to the specifically exemplified sequences. For example, one of skill will appreciate that the invention also provides, e.g., many related sequences with the functions described herein, e.g., polynucleotides and polypeptides encoding conservative variants of a polymerase of Table A or any other specifically listed polymerase herein. Combinations of any of the mutations noted herein or combinations of any of the mutations herein in combination with those noted in other

available references relating to improved polymerases, such as Hanzel et WO 2007/076057 POLYMERASES FOR NUCLEOTIDE ANALOG INCORPORATION; Rank et al. PCT/US2007/022459 POLYMERASE ENZYMES AND REAGENTS FOR ENHANCED NUCLEIC ACID SEQUENCING; Hanzel et al. WO 2007/075987 ACTIVE SURFACE COUPLED POLYMERASES; and Hanzel et al. WO 2007/076057 PROTEIN ENGINEERING STRATEGIES TO OPTIMIZE ACTIVITY OF SURFACE ATTACHED PROTEINS are also features of the invention

[0242] Accordingly, the invention provides a variety of polypeptides (polymerases) and polynucleotides (nucleic acids that encode polymerases). Example polynucleotides of the invention include, e.g., a polynucleotide encoding a polymerase as set forth in Table A or a polynucleotide that is complementary to or that encodes a polynucleotide sequence thereof (e.g., where the given sequence is a DNA, an RNA is one example of a sequence that encodes the DNA, e.g., via reverse transcription). A polynucleotide of the invention also optionally includes any polynucleotide that encodes a polymerase of Table A. Because of the degeneracy of the genetic code, many polynucleotides equivalently encode a given polymerase sequence. Similarly, an artificial or recombinant nucleic acid that hybridizes to a polynucleotide indicated above under highly stringent conditions over substantially the entire length of the nucleic acid (and is other than a naturally occurring polynucleotide) is a polynucleotide of the invention. In one embodiment, a composition includes a polypeptide of the invention and an excipient (e.g., buffer, water, pharmaceutically acceptable excipient, etc.). The invention also provides an antibody or antisera specifically immunoreactive with a polypeptide of the invention (e.g., that specifically recognizes a feature of the polymerase that confers increased branching, delayed translocation or increased nucleotide analog residence time.

[0243] In certain embodiments, a vector (e.g., a plasmid, a cosmid, a phage, a virus, etc.) comprises a polynucleotide of the invention. In one embodiment, the vector is an expression vector. In another embodiment, the expression vector includes a promoter operably linked to one or more of the polynucleotides of the invention. In another embodiment, a cell comprises a vector that includes a polynucleotide of the invention.

[0244] One of skill will also appreciate that many variants of the disclosed sequences are included in the invention. For example, conservative variations of the disclosed sequences that yield a functionally similar sequence are included in the invention. Variants of the nucleic acid polynucleotide sequences, wherein the variants hybridize to at least one disclosed sequence, are considered to be included in the invention. Unique subsequences of the sequences disclosed herein, as determined by, e.g., standard sequence comparison techniques, are also included in the invention.

Conservative Variations

[0245] Owing to the degeneracy of the genetic code, "silent substitutions" (*i.e.*, substitutions in a nucleic acid sequence which do not result in an alteration in an encoded polypeptide) are an implied feature of *every* nucleic acid sequence that encodes an amino acid sequence. Similarly, "conservative amino acid substitutions," where one or a limited number of amino acids in an amino acid sequence are substituted with different amino acids with highly similar properties, are also readily identified as being highly similar to a disclosed construct. Such conservative variations of each disclosed sequence are a feature of the present invention.

[0246] "Conservative variations" of a particular nucleic acid sequence refers to those nucleic acids which encode identical or essentially identical amino acid sequences, or, where the nucleic acid does not encode an amino acid sequence, to essentially identical sequences. One of skill will recognize that individual substitutions, deletions or additions which alter, add or delete a single amino acid or a small percentage of amino acids (typically less than 5%, more typically less than 4%, 2% or 1%) in an encoded sequence are "conservatively modified variations" where the alterations result in the deletion of an amino acid, addition of an amino acid, or substitution of an amino acid with a chemically similar amino acid, while retaining the relevant mutational feature (for example, the conservative substitution can be of a residue distal to the active site region, or distal to an interdomain stability region). Thus, "conservative variations" of a listed polypeptide sequence of the present invention include substitutions of a small percentage, typically less than 5%, more typically less than 2% or 1%, of the amino acids of the polypeptide sequence, with an amino acid of the same conservative substitution group. Finally, the addition of sequences which do not alter the encoded activity of a nucleic acid

molecule, such as the addition of a non-functional or tagging sequence (introns in the nucleic acid, poly His or similar sequences in the encoded polypeptide, etc.), is a conservative variation of the basic nucleic acid or polypeptide.

[0247] Conservative substitution tables providing functionally similar amino acids are well known in the art, where one amino acid residue is substituted for another amino acid residue having similar chemical properties (e.g., aromatic side chains or positively charged side chains), and therefore does not substantially change the functional properties of the polypeptide molecule. The following sets forth example groups that contain natural amino acids of like chemical properties, where substitutions within a group is a "conservative substitution".

TABLE C
Conservative Amino Acid Substitutions

Nonpolar and/or Aliphatic Side Chains	Polar, Uncharged Side Chains	Aromatic Side Chains	Positively Charged Side Chains	Negatively Charged Side Chains
Glycine	Serine			
Alanine	Threonine	Phenylalanine	Lysine	Aspartate
Valine	Cysteine	Tyrosine	Arginine	Glutamate
Leucine	Methionine	Tryptophan	Histidine	
Isoleucine	Asparagine			
Proline	Glutamine			

Nucleic Acid Hybridization

[0248] Comparative hybridization can be used to identify nucleic acids of the invention, including conservative variations of nucleic acids of the invention. In addition, target nucleic acids which hybridize to a nucleic acid encoding a polymerase of Table A or any other specifically listed polymerase herein under high, ultra-high and ultra-ultra high stringency conditions, where the nucleic acids encode mutations corresponding to those noted in Table A or other listed polymerases are a feature of the invention. Examples of such nucleic acids include those with one or a few silent or conservative nucleic acid substitutions as compared to a given nucleic acid sequence encoding a polymerase of Table A (or other exemplified

polymerase), where any conservative substitutions are for residues other than those noted in Table A or elsewhere as being relevant to a feature of interest (improved complex stability, decreased branch rate formation, etc.).

[0249] A test nucleic acid is said to specifically hybridize to a probe nucleic acid when it hybridizes at least 50% as well to the probe as to the perfectly matched complementary target, i.e., with a signal to noise ratio at least half as high as hybridization of the probe to the target under conditions in which the perfectly matched probe binds to the perfectly matched complementary target with a signal to noise ratio that is at least about 5x-10x as high as that observed for hybridization to any of the unmatched target nucleic acids.

[0250] Nucleic acids "hybridize" 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) provide details on the synthesis, labeling, detection and quantification of DNA and RNA, including oligonucleotides.

[0251] An example of stringent hybridization conditions for hybridization of complementary nucleic acids which have more than 100 complementary residues on a filter in a Southern or northern blot is 50% formalin with 1 mg of heparin at 42°C, with the hybridization being carried out overnight. An example of stringent wash conditions is a 0.2x SSC wash at 65°C for 15 minutes (*see,* Sambrook, *supra* for a description of SSC buffer). Often the high stringency wash is preceded by a low stringency wash to remove background probe signal. An example low stringency wash is 2x SSC at 40°C for 15 minutes. In general, a signal to noise

ratio of 5x (or higher) than that observed for an unrelated probe in the particular hybridization assay indicates detection of a specific hybridization.

[0252] "Stringent hybridization wash conditions" in the context of nucleic acid hybridization experiments such as Southern and northern hybridizations are sequence dependent, and are different under different environmental parameters. An extensive guide to the hybridization of nucleic acids is found in Tijssen (1993), *supra.* and in Hames and Higgins, 1 and 2. Stringent hybridization and wash conditions can easily be determined empirically for any test nucleic acid. For example, in determining stringent hybridization and wash conditions, the hybridization and wash conditions are gradually increased (e.g., by increasing temperature, decreasing salt concentration, increasing detergent concentration and/or increasing the concentration of organic solvents such as formalin in the hybridization or wash), until a selected set of criteria are met. For example, in highly stringent hybridization and wash conditions, the hybridization and wash conditions are gradually increased until a probe binds to a perfectly matched complementary target with a signal to noise ratio that is at least 5x as high as that observed for hybridization of the probe to an unmatched target.

[0253] "Very stringent" conditions are selected to be equal to the thermal melting point (T_m) for a particular probe. The T_m is the temperature (under defined ionic strength and pH) at which 50% of the test sequence hybridizes to a perfectly matched probe. For the purposes of the present invention, generally, "highly stringent" hybridization and wash conditions are selected to be about 5° C lower than the T_m for the specific sequence at a defined ionic strength and pH.

[0254] "Ultra high-stringency" hybridization and wash conditions are those in which the stringency of hybridization and wash conditions are increased until the signal to noise ratio for binding of the probe to the perfectly matched complementary target nucleic acid is at least 10x as high as that observed for hybridization to any of the unmatched target nucleic acids. A target nucleic acid which hybridizes to a probe under such conditions, with a signal to noise ratio of at least ½ that of the perfectly matched complementary target nucleic acid is said to bind to the probe under ultra-high stringency conditions.

[0255] Similarly, even higher levels of stringency can be determined by gradually increasing the hybridization and/or wash conditions of the relevant hybridization assay. For example, those in which the stringency of hybridization and wash conditions are increased until the signal to noise ratio for binding of the probe to the perfectly matched complementary target nucleic acid is at least 10x, 20X, 50X, 100X, or 500X or more as high as that observed for hybridization to any of the unmatched target nucleic acids. A target nucleic acid which hybridizes to a probe under such conditions, with a signal to noise ratio of at least ½ that of the perfectly matched complementary target nucleic acid is said to bind to the probe under ultra-ultra-high stringency conditions.

[0256] Nucleic acids that do not hybridize to each other under stringent conditions are still substantially identical if the polypeptides which they encode are substantially identical. This occurs, *e.g.*, when a copy of a nucleic acid is created using the maximum codon degeneracy permitted by the genetic code.

Unique Subsequence

[0257] In some aspects, the invention provides a nucleic acid that comprises a unique subsequence in a nucleic acid that encodes a polymerase of Table A. The unique subsequence may be unique as compared to a nucleic acid corresponding to, e.g., a wild type Φ29. Alignment can be performed using, e.g., BLAST set to default parameters. Any unique subsequence is useful, e.g., as a probe to identify the nucleic acids of the invention.

[0258] Similarly, the invention includes a polypeptide which comprises a unique subsequence in a polymerase of Table A. Here, the unique subsequence is unique as compared to, e.g., wild type Φ29 or previously characterized mutation thereof.

[0259] The invention also provides for target nucleic acids which hybridize under stringent conditions to a unique coding oligonucleotide which encodes a unique subsequence in a polypeptide selected from the sequences of Table A, wherein the unique subsequence is unique as compared to a polypeptide corresponding to wild type Φ29. Unique sequences are determined as noted above.

Sequence comparison, identity and homology

[0260] The terms "identical" or "percent identity," in the context of two or more nucleic acid or polypeptide sequences, refer to two or more sequences or subsequences that are the same or have a specified percentage of amino acid residues or nucleotides that are the same, when compared and aligned for maximum correspondence, as measured using one of the sequence comparison algorithms described below (or other algorithms available to persons of skill) or by visual inspection.

[0261] The phrase "substantially identical," in the context of two nucleic acids or polypeptides (*e.g.*, DNAs encoding a polymerase, or the amino acid sequence of a polymerase) refers to two or more sequences or subsequences that have at least about 60%, about 80%, about 90-95%, about 98%, about 99% or more nucleotide or amino acid residue identity, when compared and aligned for maximum correspondence, as measured using a sequence comparison algorithm or by visual inspection. Such "substantially identical" sequences are typically considered to be "homologous," without reference to actual ancestry. Preferably, the "substantial identity" exists over a region of the sequences that is at least about 50 residues in length, more preferably over a region of at least about 100 residues, and most preferably, the sequences are substantially identical over at least about 150 residues, or over the full length of the two sequences to be compared.

[0262] Proteins and/or protein sequences are "homologous" when they are derived, naturally or artificially, from a common ancestral protein or protein sequence. Similarly, nucleic acids and/or nucleic acid sequences are homologous when they are derived, naturally or artificially, from a common ancestral nucleic acid or nucleic acid sequence. Homology is generally inferred from sequence similarity between two or more nucleic acids or proteins (or sequences thereof). The precise percentage of similarity between sequences that is useful in establishing homology varies with the nucleic acid and protein at issue, but as little as 25% sequence similarity over 50, 100, 150 or more residues is routinely used to establish homology. Higher levels of sequence similarity, e.g., 30%, 40%, 50%, 60%, 70%, 80%, 90%, 95%, or 99% or more, can also be used to establish homology. Methods for determining sequence similarity percentages (e.g., BLASTP and BLASTN using default parameters) are described herein and are generally available.

[0263] For sequence comparison and homology determination, typically one sequence acts as a reference sequence to which test sequences are compared. When using a sequence comparison algorithm, test and reference sequences are input into a computer, subsequence coordinates are designated, if necessary, and sequence algorithm program parameters are designated. The sequence comparison algorithm then calculates the percent sequence identity for the test sequence(s) relative to the reference sequence, based on the designated program parameters.

[0264] Optimal alignment of sequences for comparison can be conducted, *e.g.*, by the local homology algorithm of Smith & Waterman, Adv. Appl. Math. 2:482 (1981), by the homology alignment algorithm of Needleman & Wunsch, J. Mol. Biol. 48:443 (1970), by the search for similarity method of Pearson & Lipman, Proc. Nat'l. Acad. Sci. USA 85:2444 (1988), by computerized implementations of these algorithms (GAP, BESTFIT, FASTA, and TFASTA in the Wisconsin Genetics Software Package, Genetics Computer Group, 575 Science Dr., Madison, WI), or by visual inspection (*see generally* 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).

[0265] One example of an algorithm that is suitable for determining percent sequence identity and sequence similarity is the BLAST algorithm, which is described in Altschul *et al.*, J. Mol. Biol. 215:403-410 (1990). Software for performing BLAST analyses is publicly available through the National Center for Biotechnology Information. This algorithm involves first identifying high scoring sequence pairs (HSPs) by identifying short words of length W in the query sequence, which either match or satisfy some positive-valued threshold score T when aligned with a word of the same length in a database sequence. T is referred to as the neighborhood word score threshold (Altschul *et al.*, *supra*). These initial neighborhood word hits act as seeds for initiating searches to find longer HSPs containing them. The word hits are then extended in both directions along each sequence for as far as the cumulative alignment score can be increased. Cumulative scores are calculated using, for nucleotide sequences, the parameters M (reward score for a pair of matching residues; always > 0) and N (penalty score for mismatching residues; always < 0). For amino acid sequences, a scoring matrix is used to calculate the cumulative score. Extension of the word hits in each direction are halted when:

the cumulative alignment score falls off by the quantity X from its maximum achieved value; the cumulative score goes to zero or below, due to the accumulation of one or more negative-scoring residue alignments; or the end of either sequence is reached. The BLAST algorithm parameters W, T, and X determine the sensitivity and speed of the alignment. The BLASTN program (for nucleotide sequences) uses as defaults a wordlength (W) of 11, an expectation (E) of 10, a cutoff of 100, M=5, N=-4, and a comparison of both strands. For amino acid sequences, the BLASTP program uses as defaults a wordlength (W) of 3, an expectation (E) of 10, and the BLOSUM62 scoring matrix (*see* Henikoff & Henikoff (1989) Proc. Natl. Acad. Sci. USA 89:10915).

[0266] In addition to calculating percent sequence identity, the BLAST algorithm also performs a statistical analysis of the similarity between two sequences (*see, e.g.*, Karlin & Altschul, Proc. Nat'l. Acad. Sci. USA 90:5873-5787 (1993)). One measure of similarity provided by the BLAST algorithm is the smallest sum probability (P(N)), which provides an indication of the probability by which a match between two nucleotide or amino acid sequences would occur by chance. For example, a nucleic acid is considered similar to a reference sequence if the smallest sum probability in a comparison of the test nucleic acid to the reference nucleic acid is less than about 0.1, more preferably less than about 0.01, and most preferably less than about 0.001.

[0267] 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 determining which nucleotide is incorporated at a site during a template dependent polymerization reaction, the method comprising:

incorporating the nucleotide into a nucleic acid polymer whereby signal pulses or signatures are generated from branch fraction nonincorporation events and, optionally, actual nucleotide incorporation events, for the site;

monitoring a time course of at least branch fraction signal pulses or signatures produced by the polymerization reaction; and,

assigning which nucleotide is incorporated at the site, using at least signal pulses or signatures from branch fraction nonincorporation sampling events.

2. The method of claim **1**, wherein the method comprises counting or estimating the number of redundant iterative sampling signal pulses per incorporation event, or determining an average number of redundant signal pulses per incorporation event.

3. The method of claim **1**, wherein the polymerization reaction is a high branch fraction polymerization reaction.

4. The method of claim **3**, wherein the branch fraction is at least about 80% or more.

5. The method of claim **1**, wherein the reaction comprises at least one species of metal ion, which metal ion increases the frequency of branch fraction nonincorporation events in the reaction.

6. The method of claim **5**, wherein the at least one species of metal ion is selected from Mg^{++}, Mn^{++}, Zn^{++}, Co^{++}, Ca^{++}, Fe^{++}, Cr^{++} and Sr^{++}.

7. The method of claim **5**, wherein the reaction comprises Mg^{++} and Mn^{++}.

8. The method of claim **7**, wherein the concentration of Mg^{++} is higher than the concentration of Mn^{++}.

9. The method of claim **1**, wherein the reaction comprises a Φ29, B103, GA-1, PZA, Φ15, BS32, M2Y, Nf, G1, Cp-1, PRD1, PZE, SF5, Cp-5, Cp-7, PR4, PR5, PR722, L17, T4 or T7 DNA polymerase, or a modified recombinant DNA polymerase thereof.

10. The method of claim **9**, wherein the modified recombinant polymerase exhibits a higher branching fraction as compared to a corresponding wild-type polymerase.

11. The method of claim **9**, wherein the modified recombinant polymerase comprises an exonuclease rate that is about 10% to 50% as compared to its polymerization rate.

12. The method of claim **1**, wherein the branch fraction nonincorporation events comprise iterative sampling of labeled unincorporatable nucleotide analogs.

13. The method of claim **12**, wherein actual nucleotide incorporation events comprise incorporation of unlabeled nucleotides.

14. The method of claim **12**, wherein actual nucleotide incorporation events comprise incorporation of nucleotides that are differentially labeled as compared to the unincorporatable nucleotides.

15. The method of claim **12**, wherein the labeled unincorporatable nucleotide analogs comprise a link between an alpha and beta phosphate group that is not hydrolyzable by a polymerase enzyme.

16. The method of claim **1**, wherein a polymerase enzyme, polymerase reaction conditions, and/or polymerase reaction substrates are selected such that the polymerization reaction exhibits two kinetically observable steps within an observable phase of the polymerase reaction.

17. The method of claim **16**, wherein the two kinetically observable steps are each steps which proceed in a bright phase.

18. The method of claim **16**, wherein the two kinetically observable steps are each steps which proceed in a dark phase.

19. The method of claim **16**, wherein the polymerase enzyme comprises a modified recombinant Φ29, B103, GA-1, PZA, Φ15, BS32, M2Y, Nf, G1, Cp-1, PRD1, PZE, SF5, Cp-5, Cp-7, PR4, PR5, PR722, L17, T4 or T7 DNA polymerase.

20. The method of claim **16**, wherein the polymerase reaction conditions comprise one or more of: a selected metal cofactor concentration, a selected pH, a selected temperature, an enzyme activity modulator, D_2O, an organic solvent, and a buffer.

21. The method of claim **1**, wherein the branch fraction nonincorporation events comprise noncognate branch fraction nonincorporation events.

22. The method of claim **21**, wherein the branch fraction signal pulses or signatures are generated from noncognate branching events of a nucleotide analog comprising a base selected from: guanine and thymine.

23. The method of claim **1**, wherein the reaction is reacted in a DNA sequencing system.

24. The method of claim **23**, wherein the DNA sequencing system comprises a zero mode waveguide or nanohole.

25. The method of claim **1**, wherein assigning the nucleotide comprises applying a statistical model to the signal pulses or signatures generated from branch fraction nonincorporation events, signal pulses generated from actual nucleotide incorporation events, or both, which statistical model assigns a likelihood that the signal pulses or signatures correspond to an incorporation event.

26. The method of claim **1**, further comprising:

performing an additional template dependent polymerization reaction under high processivity reaction conditions;

monitoring a second time course of signal pulses or signatures produced by the additional polymerization reaction; and,

compiling sequencing information derived from the second time course of signal pulses or signatures with sequencing information derived from the time course of branch fraction signal pulses or signatures.

27. A multi-modal sequencing method, comprising:

performing a first template dependent sequencing reaction in a first mode comprising a first set of reaction conditions;

collecting initial sequencing information produced by the first sequencing reaction;

performing a second sequencing reaction of the template, or a copy thereof, in a second mode comprising a second set of reaction conditions;

collecting additional sequencing information produced by the second sequencing reaction; and,

compiling the initial and additional sequencing information to provide a sequence of at least a portion of the template.

28. The method of claim **27**, wherein the second sequencing reaction is produced by altering one or more reaction conditions of the first sequencing reaction, wherein the initial and additional sequencing information are collected in real time.

29. The method of claim **28**, wherein altering one or more reaction conditions comprises adding one or more polymerase cofactors to the first sequencing reaction.

30. The method of claim **28**, wherein altering one or more reaction conditions comprises adding Mn^{++} or Mg^{++} to the first sequencing reaction.

31. The method of claim **27**, wherein both the first and second sequencing reactions comprise single template molecule sequencing reactions, and wherein the first mode produces a higher branch fraction as compared to the second mode.

32. The method of claim **27**, wherein both the first and second sequencing reactions comprise single template molecule sequencing reactions, and wherein the second mode produces longer read lengths as compared to the first mode.

33. The method of claim **27**, wherein the template is a single circular template molecule, and the method comprises switching between the first and second modes.

34. A method of determining which of two or more labeled nucleotides is incorporated at a site during a template-dependent polymerization reaction, the method comprising:

incorporating the nucleotide into a nucleic acid polymer produced by the polymerization reaction, whereby signal pulses or signatures are generated;

monitoring the pulses or signatures; and,

using the presence of multiple pulses corresponding to the nucleotide, or identical molecules thereof, to assign which labeled nucleotide is incorporated at the site.

35. The method of claim **34**, wherein the multiple pulses comprise 2 to 20 pulses.

36. The method of claim **34**, wherein the multiple pulses are generated from branch fraction nonincorporation events.

37. The method of claim **36**, wherein the branch fraction nonincorporation events are induced by a sequencing composition that comprises at least one species of metal ion.

38. The method of claim **37**, wherein the metal ion is selected from: Mg^{++}, Mn^{++}, Zn^{++}, Co^{++}, Ca^{++}, Fe^{++}, Cr^{++} and Sr^{++}.

39. The method of claim **34**, wherein using the presence of multiple pulses comprises distinguishing incorporation and nonincorporation signals to assign which labeled nucleotide incorporated at the site.

40. A nucleic acid sequencing system that, during operation of the system, sequences a nucleic acid, the system comprising:

a signal detector, which detector detects at least signal pulses or signatures from branch fraction nonincorporation events during sequencing of a template nucleic acid;

system instructions or software that assign a sequence based upon detection of at least signal pulses or signatures from branch fraction nonincorporation events; and,

a user output module that displays the sequence to the user.

41. The nucleic acid sequencing system of claim **40**, wherein the signal pulses or signatures from branch fraction nonincorporation events are generated during a first sequencing mode, wherein the signal detector detects signal pulses generated during a low branch fraction second sequencing mode, and wherein the system instructions assign a sequence based upon detection of signal pulses or signatures from the first and second sequencing modes.

42. The nucleic acid sequencing system of claim **40**, wherein the system comprises a zero-mode waveguide or nanohole proximal to the signal detector, and wherein, during operation of the system, a sequencing reaction is contained by the zero-mode waveguide or nanohole.

43. A composition comprising a modified recombinant nucleic acid polymerase, which modified recombinant polymerase exhibits an altered property selected from:

a) an increased branching fraction during a polymerization by the polymerase;

b) an altered translocation property of the polymerase during a polymerization reaction; and,

c) a combination of any of (a)-(b),

wherein the altered property is altered as compared to a corresponding wild-type polymerase.

44. The composition of claim **43**, wherein the modified recombinant polymerase is a modified recombinant Φ29, B103, GA-1, PZA, Φ15, BS32, M2Y, Nf, G1, Cp-1, PRD1, PZE, SF5, Cp-5, Cp-7, PR4, PR5, PR722, L17, T4, or T7 polymerase.

45. The composition of claim **43**, wherein the modified recombinant polymerase that exhibits an increased branching fraction comprises at least one amino acid substitution or deletion or combination of substitutions or deletions selected from: N62D and Y454A; D362S; Y259H; F237Y; L381I; Y369H; H461Y; A377G; K138Q; H461D; A377S; N62D and K371Q; V118L; and K124R; wherein numbering of the residues is relative to a wild-type Φ29 polymerase of SEQ ID NO: 3.

46. The composition of claim **43**, wherein the modified recombinant polymerase that exhibits an increased branching fraction exhibits a branching fraction that is at least 50% greater than the branching fraction of a wild-type Φ29 polymerase of SEQ ID NO: 3.

47. The composition of claim **43**, wherein the modified recombinant polymerase that exhibits an increased branching fraction exhibits a branching fraction that is at least 100% greater than the branching fraction of a wild-type Φ29 polymerase of SEQ ID NO: 3.

48. The composition of claim **43**, wherein the modified recombinant polymerase that exhibits an increased branching fraction exhibits a branching fraction that is at least 200% greater than the branching fraction of a wild-type Φ29 polymerase of SEQ ID NO: 3.

49. The composition of claim **43**, wherein the modified recombinant DNA polymerase that exhibits an increased branching fraction exhibits increased exonuclease activity as compared to the corresponding wild type polymerase.

50. The composition of claim **49**, wherein the modified recombinant polymerase comprises an exonuclease rate that is about 10% to 50% as compared to its polymerization rate.

51. The composition of claim **43**, wherein the modified recombinant polymerase that exhibits an altered translocation property comprises a fusion protein, which fusion protein comprises at least a subsequence of the parental polymerase and at least one heterologous polypeptide sequence.

52. The composition of claim **51**, wherein the fusion protein comprises the at least one heterologous polypeptide sequence fused at or near the c-terminal end of the parental polymerase.

53. The composition of claim **52**, wherein the wild-type polymerase is a Φ29 polymerase.

54. The composition of claim **51**, wherein the heterologous polypeptide sequence comprises at least one charged amino acid.

55. The composition of claim **54**, wherein the at least one charged amino acid is histidine.

56. The composition of claim **51**, wherein the fusion protein comprises a linker between the at least a subsequence of the parental polymerase and the heterologous polypeptide sequence.

57. The composition of claim **56**, wherein the linker comprises a Ser3Gly linker.

58. The composition of claim **43**, wherein the modified recombinant polymerase that exhibits an altered translocation property comprises at least one amino acid substitution or deletion or combination of substitutions or deletions selected from: Asp570Lys; Asp570Ala; Asn313Lys; Asn313Ala; Gln303Lys; Gln303Ala; Gly532Ser; Met533delet; Cys530delet; Met533delet and Cys530delet; Gly532delet; Ala531Gly; Thr573Lys; Thr573Ala; Asn396Lys;

Thr571Lys; Thr571Ala; Thr534Lys; Thr534Ala; Asp535Lys; Asp535Ala; Lys529Ala; and Lys529Asn; wherein numbering of the residue positions is relative to a wild-type Φ29 polymerase of SEQ ID NO: 3.

59. The composition of claim **43**, wherein the modified recombinant DNA polymerase that exhibits an altered translocation property exhibits a delay in translocation as compared to the corresponding wild-type polymerase.

60. The composition of claim **59**, wherein the delay in translocation is at least about 2.5 times greater than that of the corresponding wild-type polymerase.

61. The composition of claim **59**, wherein the delay in translocation is at least about 10 times greater than that of the corresponding wild-type polymerase.

62. The composition of claim **59**, wherein the delay in translocation is at least about 15 times greater than that of the corresponding wild-type polymerase.

63. The composition of claim **43**, wherein the modified recombinant polymerase exhibits an increased nucleotide or nucleotide analog residence time as compared to the corresponding wild-type polymerase.

64. The composition of claim **43**, wherein the modified recombinant polymerase exhibits increased processivity as compared to the corresponding wild-type polymerase.

65. The composition of claim **43**, further comprising a phosphate-labeled nucleotide analog and a DNA template, wherein the modified recombinant polymerase incorporates the nucleotide analog into a copy nucleic acid in response to the DNA template.

66. The composition of claim **43**, wherein the composition is present in a DNA sequencing system.

67. The composition of claim **66**, wherein the DNA sequencing system comprises at least one zero mode waveguide or nanohole.

68. The composition of claim **43**, wherein the modified recombinant polymerase is immobilized on a surface.

69. A method of sequencing a nucleic acid template, the method comprising:

a) providing a reaction mixture comprising:

i) the nucleic acid template;

ii) a replication initiating moiety that complexes with or is integral to the template;

iii) the modified recombinant nucleic acid polymerase of claim **43**, wherein the polymerase is capable of replicating at least a portion of the template using the moiety in a template-dependent polymerization reaction; and

iv) one or more nucleotides and/or nucleotide analogs;

b) subjecting the reaction mixture to a polymerization reaction in which the modified recombinant polymerase replicates at least a portion of the template in a template-dependent manner, whereby the one or more nucleotides and/or nucleotide analogs are incorporated into the resulting copy nucleic acid; and

c) identifying a time sequence of incorporation of the one or more nucleotides and/or nucleotide analogs into the resulting copy nucleic acid.

70. The method of claim **69**, wherein the modified recombinant polymerase exhibits increased processivity relative to the wild-type polymerase.

71. The method of claim **69**, wherein identifying the time sequence of incorporation comprises observing more than one signal pulse per nucleotide incorporation event.

72. The method of claim **69**, wherein the polymerization reaction is performed in a zero mode waveguide or nanohole.

73. A method of making a nucleic acid, the method comprising:

a) providing a reaction mixture comprising:

i) a template,

ii) a replication initiating moiety that complexes with or is integral to the template,

iii) the modified recombinant nucleic acid polymerase of claim **43**, wherein the polymerase is capable of replicating at least a portion of the template using the moiety in a template-dependent polymerase reaction, and

iv) one or more nucleotides and/or nucleotide analogs; and

b) reacting the mixture such that the polymerase replicates at least a portion of the template in a template-dependent manner, whereby the one or more nucleotides and/or nucleotide analogs are incorporated into the resulting nucleic acid.

74. The method of claim **73**, the method comprising detecting incorporation of at least one of the nucleotides and/or nucleotide analogs.

75. The method of claim **74**, wherein the detection comprises observing more than one signal pulse per nucleotide incorporation event.

76. The method of claim **73**, wherein the mixture is reacted in a zero mode waveguide or nanohole.

77. The method of claim **73**, wherein the modified recombinant polymerase exhibits an increased nucleotide residence time as compared to the parental polymerase.

78. The method of claim **73**, wherein the modified recombinant polymerase exhibits increased processivity as compared to the parental polymerase.

79. A method of making a modified recombinant DNA polymerase, the method comprising:

mutating a polymerase of interest; and

selecting resulting modified polymerases for a property selected from

i) increased branching fraction; and

ii) altered translocation.

80. The method of claim **79**, wherein mutating the polymerase of interest comprises structurally modeling the polymerase to identify a feature that may affect branch fraction or altered translocation.

81. The method of claim **79**, wherein mutating the polymerase of interest comprises making a library of modified recombinant polymerases, and wherein selecting the modified polymerases comprises screening the library to identify at least one member exhibiting the property.

82. The method of claim **79**, wherein the polymerase of interest is a Φ29, B103, GA-1, PZA, Φ15, BS32, M2Y, Nf, G1, Cp-1, PRD1, PZE, SF5, Cp-5, Cp-7, PR4, PR5, PR722, L17, T4 or T7 polymerase.

83. The method of claim **79**, wherein the resulting modified recombinant polymerase that exhibits the property also exhibits an increased nucleotide analog residence time as compared to a corresponding wild-type polymerase.

84. The method of claim **79**, wherein the resulting modified recombinant polymerase that exhibits the property also exhibits increased processivity as compared to a corresponding wild-type polymerase.

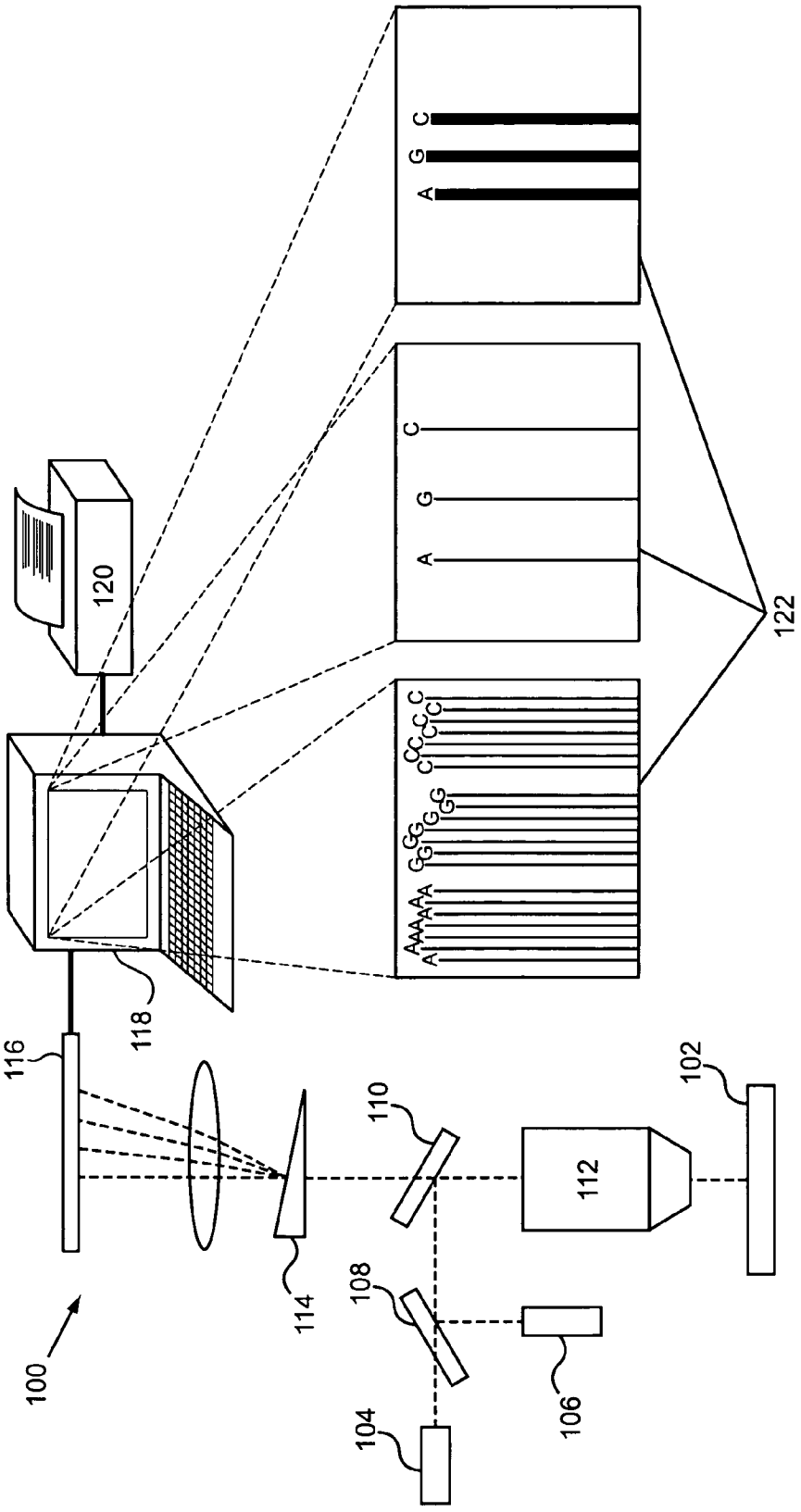

A
G
C
A
G
C
AAAAAAA
GGGGGGG
CCCCCCC
120
122
118
116
114
110
112
102
108
104
106
100

Fig. 1

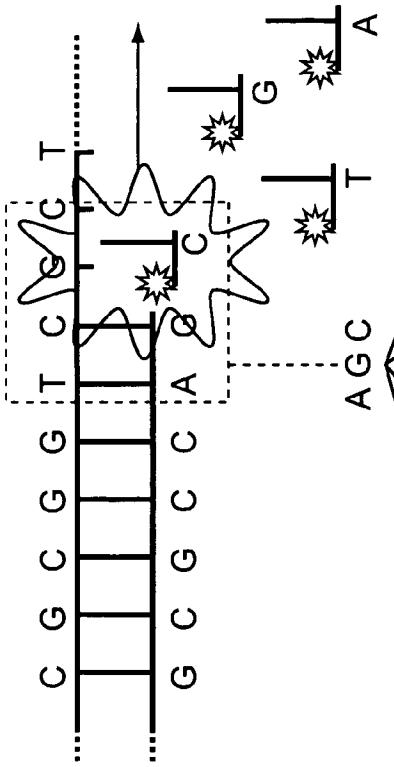

C G C G G G T C C T
G C G C C C A G C
A
G
T

Fig. 2A

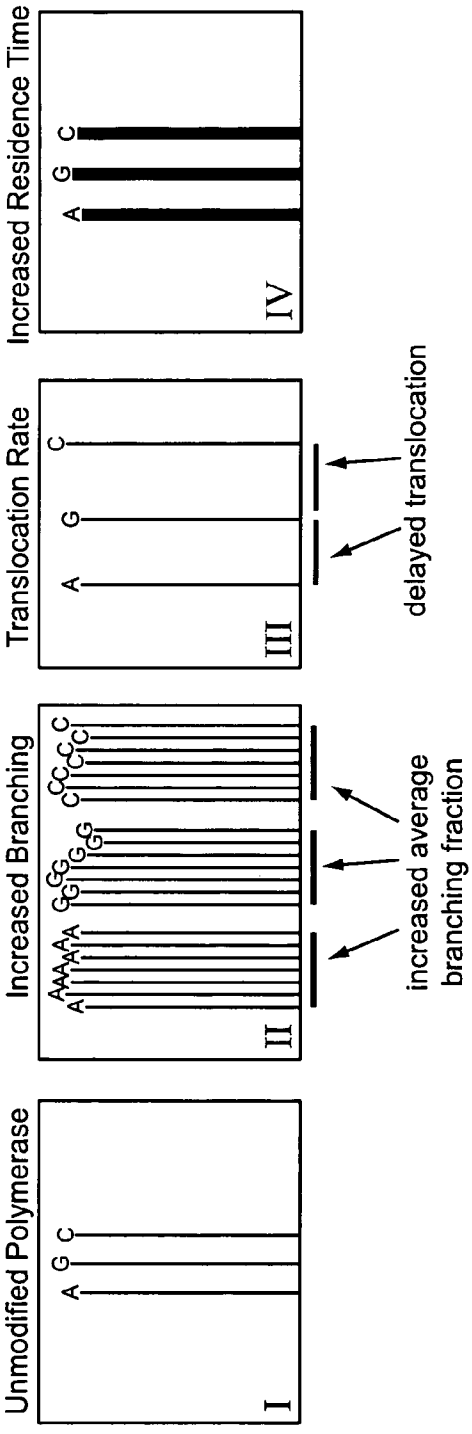

Unmodified Polymerase
I
A G C
Increased Branching
II
Decreased Translocation Rate
III
A G C
Increased Residence Time
IV
A G C
increased average branching fraction
delayed translocation

Fig. 2B

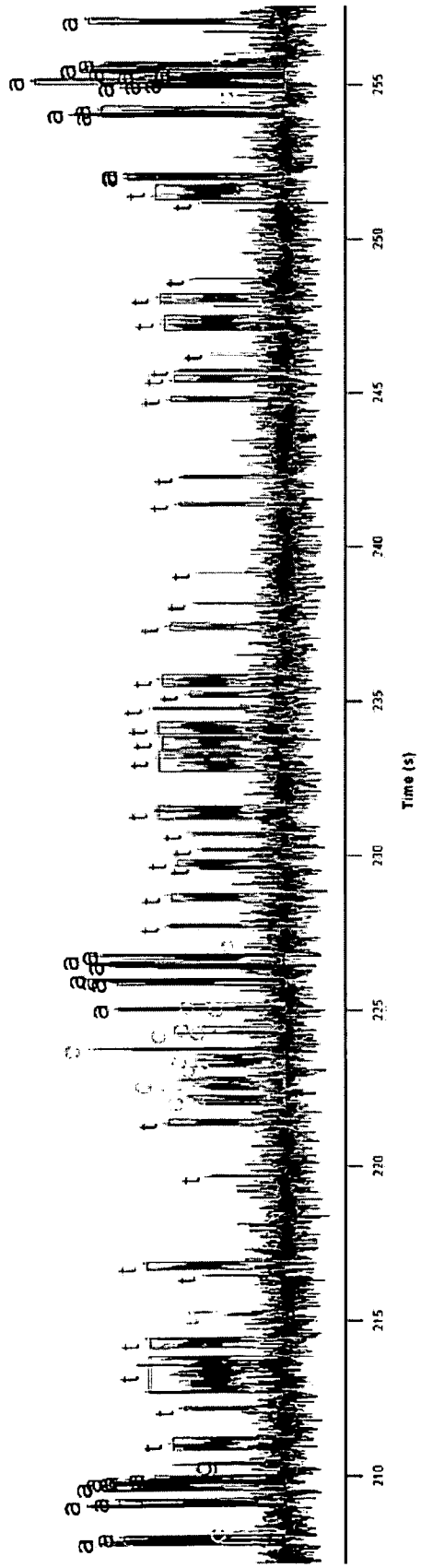


Fig. 3

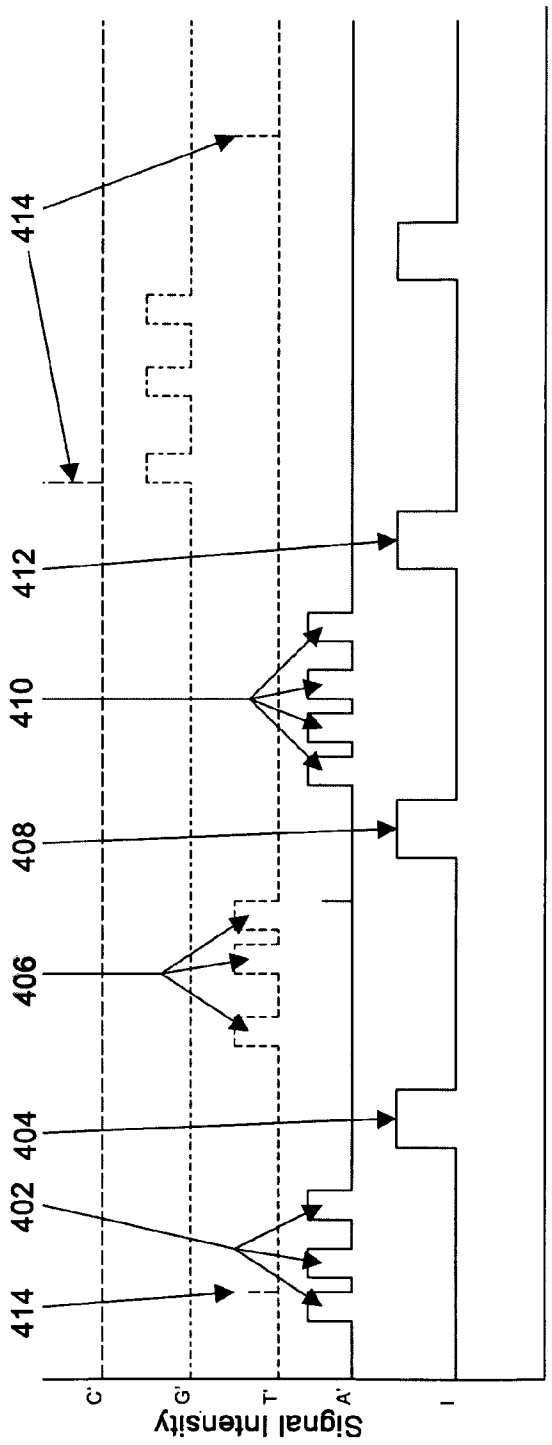


Fig. 4

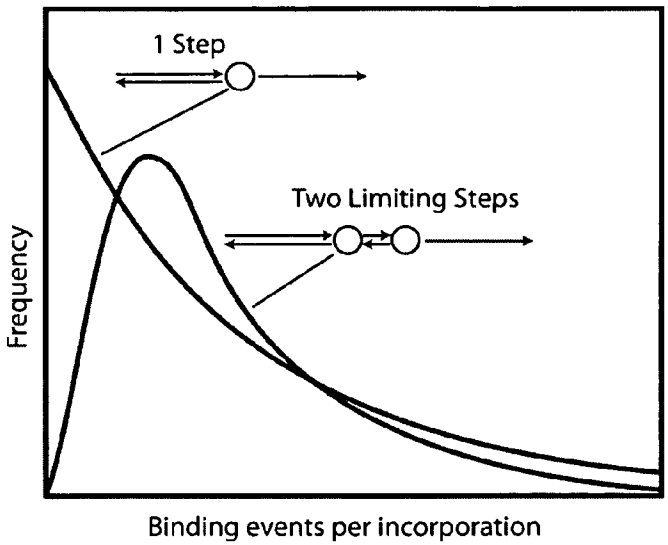

1 Step
Two Limiting Steps
Frequency
Binding events per incorporation

Fig. 5

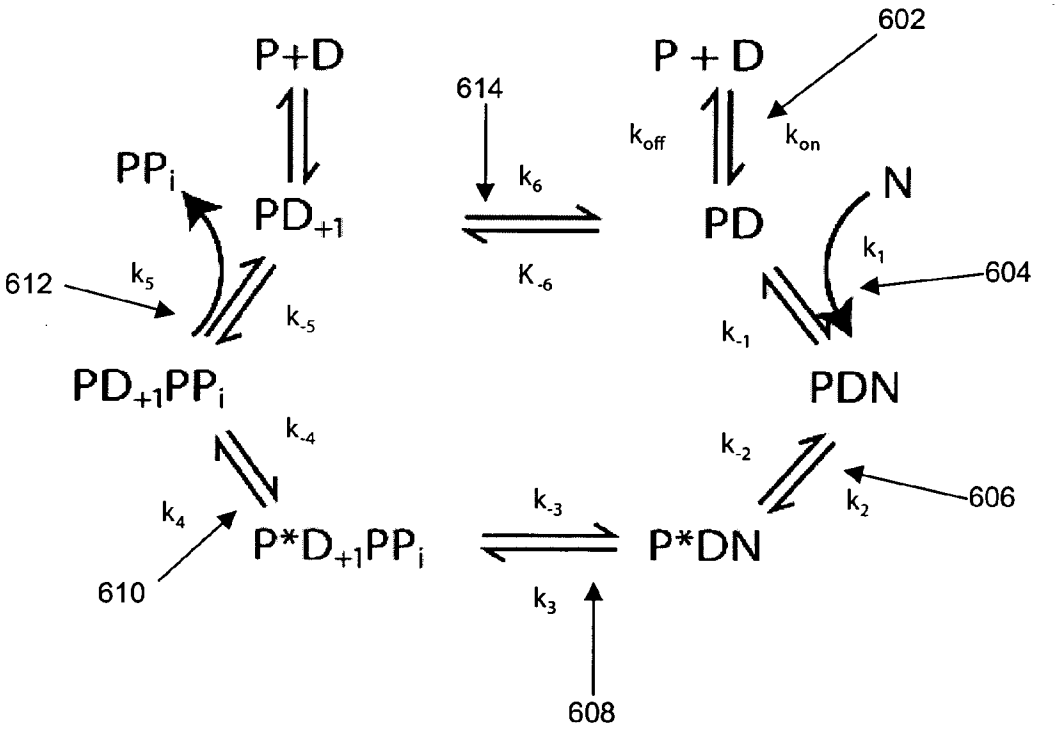

P+D
PP$_i$
PD$_{+1}$
612
k_5
k_{-5}
PD$_{+1}$PP$_i$
k_{-4}
k_4
610
P*D$_{+1}$PP$_i$
614
k_6
K_{-6}
P + D
602
k_{off}
k_{on}
N
PD
k_1
604
k_{-1}
PDN
k_{-2}
k_2
606
k_{-3}
P*DN
k_3
608

Fig. 6

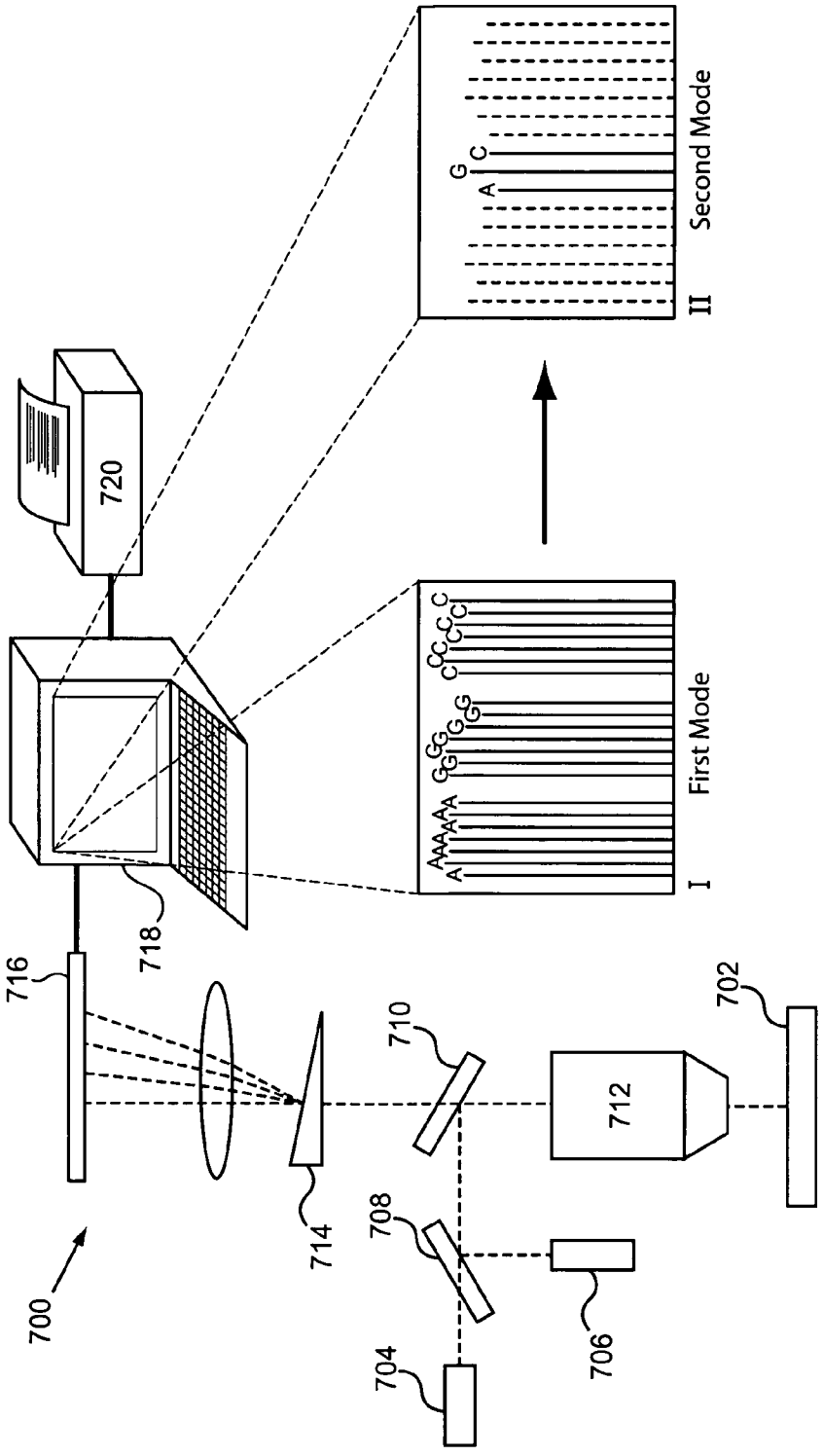

720
718
716
714
710
712
702
708
706
704
700
I
First Mode
II
Second Mode

Fig. 7

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

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
17 June 2010 (17.06.2010)

PCT

(10) International Publication Number
WO 2010/068289 A2

(51) International Patent Classification:
C12Q 1/68 (2006.01) *C12N 15/10* (2006.01)

(21) International Application Number:
PCT/US2009/006517

(22) International Filing Date:
10 December 2009 (10.12.2009)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
61/201,551 11 December 2008 (11.12.2008) US
61/180,350 21 May 2009 (21.05.2009) US
61/186,661 12 June 2009 (12.06.2009) US

(71) Applicant *(for all designated States except US)*: **PACIFIC BIOSCIENCES OF CALIFORNIA, INC.** [US/US]; 1505 Adams Drive, Menlo Park, CA 94025 (US).

(72) Inventors; and
(75) Inventors/Applicants *(for US only)*: **FLUSBERG, Benjamin** [US/US]; 533 Thain Way, Palo Alto, CA 94306 (US). **TURNER, Stephen** [US/US]; 356 Stanford Avenue, Menlo Park, CA 94025 (US). **LEE, Jessica** [US/US]; 10856 Linda Vista Drive, Cupertino, CA 95014 (US). **JIA, Lei** [CN/CN]; 3415 Janice Way, Palo Alto, CA 94303 (US). **KORLACH, Jonas** [DE/US]; 36100 Toulouse Street, Newark, CA 94560 (US). **SORENSON, Jon** [US/US]; 1725 Nason Street, Alameda, CA 94501 (US). **WEBSTER, Dale** [US/US]; 2301 Ticonderoga Drive, San Mateo, CA 94402 (US). **LYLE, John** [US/US]; 3 Avocet Drive #203, Redwood Shores, CA 94065 (US). **TRAVERS, Kevin** [US/US]; 3385 Gonzaga Place, Menlo Park, CA 95051 (US). **HANES, Jeremiah** [US/US]; 3243 Spring Street, Redwood City, CA 94063 (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, 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**: CLASSIFICATION OF NUCLEIC ACID TEMPLATES

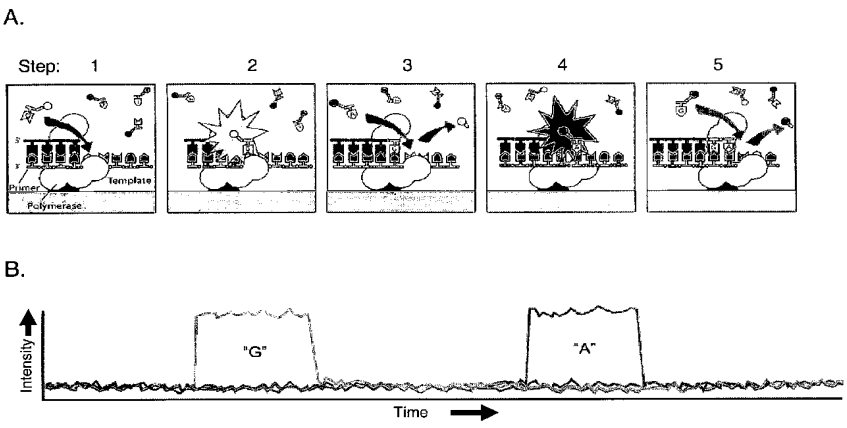


Figure 1

(57) **Abstract**: Methods, compositions, and systems are provided for characterization of modified nucleic acids. In certain preferred embodiments, single molecule sequencing methods are provided for identification of modified nucleotides within nucleic acid sequences. Modifications detectable by the methods provided herein include chemically modified bases, enzymatically modified bases, abasic sites, non-natural bases, secondary structures, and agents bound to a template nucleic acid.

WO 2010/068289 A2

CLASSIFICATION OF NUCLEIC ACID TEMPLATES

CROSS-REFERENCE TO RELATED APPLICATIONS

[0001] This application claims the benefit of Provisional U.S. Patent Application No. 61/201,551, filed December 11, 2008; Provisional U.S. Patent Application No. 61/180,350, filed May 21, 2009; and Provisional U.S. Patent Application No. 61/186,661, filed June 12, 2009, the full disclosures of which are hereby incorporated herein by reference in their entireties for all purposes.

BACKGROUND OF THE INVENTION

[0002] Assays for analysis of biological processes are exploited for a variety of desired applications. For example, monitoring the activity of key biological pathways can lead to a better understanding of the functioning of those systems as well as those factors that might disrupt the proper functioning of those systems. In fact, various different disease states caused by operation or disruption of specific biological pathways are the focus of much medical research. By understanding these pathways, one can model approaches for affecting them to prevent the onset of the disease or mitigate its effects once manifested.

A stereotypical example of the exploitation of biological process monitoring is in the area of pharmaceutical research and development. In particular, therapeutically relevant biological pathways, or individual steps or subsets of individual steps in those pathways, are often reproduced or modeled in *in vitro* systems to facilitate analysis. By observing the progress of these steps or whole pathways in the presence and absence of potential therapeutic compositions, e.g., pharmaceutical compounds or other materials, one can identify the ability of those compositions to affect the *in vitro* system, and potentially beneficially affect an organism in which the pathway is functioning in a detrimental way. By way of specific example, reversible methylation of the 5' position of cytosine by methyltransferases is one of the most widely studied epigenetic modifications. In mammals, 5-methylcytosine (5-MeC) frequently occurs at CpG dinucleotides, which often cluster in regions called CpG islands that are at or near transcription start sites. Methylation of cytosine in CpG islands can interfere with transcription factor binding and is associated with transcription repression and gene regulation. In addition, DNA methylation is known to be essential for mammalian development and has been associated with cancer and other disease processes. Recently, a new 5-hydroxymethylcytosine epigenetic marker has been identified in certain cell types in the brain, suggesting that it plays a role in

epigenetic control of neuronal function (S. Kriaucionis, et al., *Science* **2009**, 324(5929): 929-30, incorporated herein by reference in its entirety for all purposes). Further information on cytosine methylation and its impact on gene regulation, development, and disease processes is provided in the art, e.g., in A. Bird, *Genes Dev* **2002**, *16*, 6; M. Gardiner-Garden, et al., *J Mol Biol* **1987**, *196*, 261; S. Saxonov, et al., *Proc Natl Acad Sci U S A* **2006**, *103*, 1412; R. Jaenisch, et al., *Nat Genet* **2003**, *33 Suppl*, 245; E. Li, et al., *Cell* **1992**, *69*, 915; A. Razin, et al., *Hum Mol Genet* **1995**, *4 Spec No*, 1751; P. A. Jones, et al., *Nat Rev Genet* **2002**, *3*, 415; P. A. Jones, et al., *Nat Genet* **1999**, *21*, 163; and K. D. Robertson, *Nat Rev Genet* **2005**, *6*, 597, all of which are incorporated herein by reference in their entireties for all purposes.

[0004] In contrast to determining a human genome, mapping of the human methylome is a more complex task because the methylation status differs between tissue types, changes with age, and is altered by environmental factors (P. A. Jones, et al., *Cancer Res* **2005**, *65*, 11241, incorporated herein by reference in its entirety for all purposes). Comprehensive, high-resolution determination of genome-wide methylation patterns from a given sample has been challenging due to the sample preparation demands and short read lengths characteristic of current DNA sequencing technologies (K. R. Pomraning, et al., *Methods* **2009**, *47*, 142, incorporated herein by reference in its entirety for all purposes).

[0005] Bisulfite sequencing is the current method of choice for single-nucleotide resolution methylation profiling (S. Beck, et al., *Trends Genet* **2008**, *24*, 231; and S. J. Cokus, et al., *Nature* **2008**, 452, 215, the disclosures of which are incorporated herein by reference in their entireties for all purposes). Treatment of DNA with bisulfite converts unmethylated cytosine, but not 5-MeC, to uracil (M. Frommer, et al., *Proc Natl Acad Sci U S A* **1992**, *89*, 1827, incorporated herein by reference in its entirety for all purposes). The DNA is then amplified (which converts all uracils into thymines) and subsequently analyzed with various methods, including microarray-based techniques (R. S. Gitan, et al., *Genome Res* **2002**, *12*, 158, incorporated herein by reference in its entirety for all purposes) or 2^{nd}-generation sequencing (K. H. Taylor, et al., *Cancer Res* **2007**, *67*, 8511 ; and R. Lister, et al., *Cell* **2008**, *133*, 523, both incorporated herein by reference in their entireties for all purposes). While bisulfite-based techniques have greatly advanced the analysis of methylated DNA, they also have several drawbacks. First, bisulfite sequencing requires a significant amount of sample preparation time (K. R. Pomraning, et al., *supra*). Second, the harsh reaction conditions necessary for complete conversion of unmethylated cytosine to uracil lead to degradation of DNA (C. Grunau, et al., *Nucleic Acids Res* **2001**, *29*, E65, incorporated herein by reference in its entirety for all purposes), and thus necessitate large starting amounts of the sample, which can be problematic for some applications.

[0006] Furthermore, because bisulfite sequencing relies on either microarray or 2nd-generation DNA sequencing technologies for its readout of methylation status, it also suffers from the same limitations as do these methodologies. For array-based procedures, the reduction in sequence complexity caused by bisulfite conversion makes it difficult to design enough unique probes for genome-wide profiling (S. Beck, et al., *supra*). Most 2nd-generation DNA sequencing techniques employ short reads and thus have difficulties aligning to highly repetitive genomic regions (K. R. Pomraning, et al., *supra*). This is especially problematic, since many CpG islands reside in such regions. Given these limitations, bisulfite sequencing is also not well suited for *de novo* methylation profiling (S. Beck, et al., *supra*).

[0007] In another widely used technique, methylated DNA immunoprecipitation (MeDIP), an antibody against 5-MeC is used to enrich for methylated DNA sequences (M. Weber, et al., *Nat Genet* **2005**, *37*, 853, incorporated herein by reference in its entirety for all purposes). MeDIP has many advantageous attributes for genome-wide assessment of methylation status, but it does not offer as high base resolution as bisulfite treatment-based methods. In addition, it is also hampered by the same limitations of current microarray and 2nd-generation sequencing technologies.

[0008] Research efforts aimed at increasing our understanding of the human methylome would benefit greatly from the development of a new methylation profiling technology that does not suffer from the limitations described above. Accordingly, there exists a need for improved techniques for detection of modifications in nucleic acid sequences, and particularly nucleic acid methylation.

[0009] Typically, modeled biological systems rely on bulk reactions that ascertain general trends of biological reactions and provide indications of how such bulk systems react to different effectors. While such systems are useful as models of bulk reactions *in vivo*, a substantial amount of information is lost in the averaging of these bulk reaction results. In particular, the activity of and effects on individual molecular complexes cannot generally be teased out of such bulk data collection strategies.

[0010] Single-molecule real-time analysis of nucleic acid synthesis has been shown to provide powerful advantages over nucleic acid synthesis monitoring that is commonly exploited in sequencing processes. In particular, by concurrently monitoring the synthesis process of nucleic acid polymerases as they work in replicating nucleic acids, one gains advantages of a system that has been perfected over millions of years of evolution. In particular, the natural DNA synthesis processes provide the ability to replicate whole genomes in extremely short periods of time, and do so with an extremely high level of fidelity to the underlying template being replicated.

[0011] The present invention is directed to a variety of different single-molecule real-time analyses for monitoring the progress and effectors of biological reactions, and in particular detecting modifications in nucleic acid sequences. For example, the present invention provides a direct methylation sequencing technology based on observing the kinetics of single polymerase molecules in real time and with high multiplex. This technique will provide for fast and economical analysis of methylation patterns, even in repetitive genomic regions.

BRIEF SUMMARY OF THE INVENTION

[0012] The present invention is generally directed to the detection of modified nucleic acid sequences, and particularly the detection of methylated bases within nucleic acid sequences using a real time direct detection of such methylated sites. The present invention is expected to have a major impact on research aiming to illuminate the role of DNA methylation in human health.

[0013] In certain aspects of the invention, methods are provided for identification of a modification in a nucleic acid molecule. In general, a template nucleic acid comprising the modification and an enzyme capable of processing the template are provided. The template nucleic acid is contacted with the enzyme, and the subsequent processing of the template by the enzyme is monitored. A change in the processing is detected, and this change is indicative of the presence of the modification in the template. Exemplary modifications that can be detected by the methods of the invention include, but are not limited to methylated bases (e.g., 5-methylcytosine, N6-methyladenosine, etc.), pseudouridine bases, 7,8-dihydro-8-oxoguanine bases, 2'-O-methyl derivative bases, nicks, apurinic sites, apyrimidic sites, pyrimidine dimers, a cis-platen crosslinking products, oxidation damage, hydrolysis damage, bulky base adducts, thymine dimers, photochemistry reaction products, interstrand crosslinking products, mismatched bases, secondary structures, and bound agents. In preferred embodiments, nucleotides or analogs thereof that are incorporated into a nascent strand synthesized by the enzyme are distinctly labeled to allow identification of a sequence of specific nucleotides or nucleotide analogs so incorporated. In certain preferred embodiments, labels are linked to nucleotides or nucleotide analogs through a phosphate group, e.g., a phosphate group other than the alpha phosphate group. As such, the labels are removed from the nucleotide or nucleotide analog upon incorporation into the nascent strand.

[0014] In some embodiments, the template nucleic acid is treated prior to processing by the enzyme, e.g., to alter the modification. The treatment may be chemical or enzymatic, and includes, e.g., glycosylase modification, bisulfite modification, DMS modification, cytosine methyltransferase modification, TET1 modification, and cytidine deaminase modification. In

some embodiments, non-natural nucleotide analogs (e.g., pyrene analogs) are incorporated into a nascent strand synthesized by the enzyme. In some embodiments, the methods comprise both treatment of the template and incorporation of non-natural nucleotide analogs into the nascent strand. In some embodiments, non-natural nucleotides are incorporated into a nascent strand in a position to pair with a modification in the template. For example, a methylated cytosine in the template can be paired with a modified guanine nucleotide analog; a template modification can pair with a non-natural nucleotide analog to form a non-natural base pair, e.g., isocytosine and isoguanine; 5-methylisocytosine and isoguanine; Im-N^O and Im-O^N; A* and T*; and 8-oxoG and adenine. In some embodiments, non-incorporatable nucleotide analogs bind the template/enzyme complex, but are not incorporated into the nascent strand, and detection of this "nonproductive" binding serves as an indication of the modification in the template. Such non-incorporatable nucleotide analogs are preferably distinctly labeled to facilitate monitoring, and optionally to distinguish such binding from incorporation of incorporatable nucleotide analogs that comprise labels.

[0015] In certain embodiments, the template nucleic acid comprises regions of internal complementarity (e.g., a double-stranded portion) and at least one single-stranded portion, and preferably the modification is located within at least one of the regions of internal complementarity. In certain embodiments, the template is a circular template. In certain embodiments, the template is a circular template comprising at least two regions of internal complementarity. In certain embodiments, the enzyme is a polymerase, such as a DNA polymerase, and RNA polymerase, a reverse transcriptase, or a derivative or variant thereof. In preferred embodiments, the enzyme is a polymerase enzyme capable of strand displacement. In specific embodiments, the enzyme is a Φ29 polymerase, optionally comprising at least one mutation at a position selected from the group consisting of K392, K422, I93, M188, K392, V399, T421, K422; S95, Y101, M102; Q99, L123, K124, T189, A190; G191, S388; P127, L384, N387, S388; and L389, Y390, and G391.

[0016] Examples of changes in the processing of the template by the enzyme that are monitored in various embodiments of the invention include, but are not limited to, kinetics, processivity, signal characteristics, error metrics, signal context, and the like. In some embodiments, a change occurs only at the modification, and in other embodiments the change occurs at one or more positions proximal to the modification, which may also include the modification position.

[0017] In certain aspects, the methods further comprise mapping the modification. In certain preferred embodiments, mapping the modification comprises analyzing a portion of the sequence read that was generated immediately prior to, during, and/or immediately after detecting the change in processing to determine a sequence complementary to the template

nucleic acid; determining the complement of the sequence complementary to the template nucleic acid; and mapping the modification at a position in the template nucleic acid that is proximal to the complement of the sequence complementary to the template nucleic acid.

[0018] In certain embodiments, a change in the processing that is indicative of the modification is a kinetic difference in the processing (e.g., detected as an alteration in one or more of interpulse duration, interpulse width, processivity, etc.) and/or a change in an error metric (e.g., accuracy, an increase in binding events that do not result in incorporation, etc.) The change in processing can be indicative of the type of modification is present in the template nucleic acid, since different types of modifications have different effects on the activity and/or fidelity of the enzyme.

[0019] In preferred embodiments, the monitoring occurs in real time during the processing of the template by the enzyme. In preferred embodiments, the template nucleic acid and the enzyme form a complex that is immobilized at a reaction site on a substrate, and in more preferred embodiments a plurality of complexes are immobilized at optically resolvable reaction sites on the substrate, wherein a single complex immobilized at one of the reaction sites is optically resolvable from any other of the complexes immobilized at any other of the reaction sites. In certain embodiments, the optically resolvable reaction sites are nanometer-scale apertures in the substrate, and can be optical confinements, such as zero-mode waveguides. In preferred embodiments, the template nucleic acid is plurality of template nucleic acids that are optically resolvable from one another during the monitoring. Preferably, the template nucleic acid is not amplified prior to contacting it with the enzyme.

[0020] In some embodiments, the modification is secondary structure in the template nucleic acid, e.g., a hairpin loop, and the change in the modification is a kinetic change, e.g., an increased interpulse duration or increased pulse width. Certain methods for identifying such a secondary structure generally comprise generating a sequence read for the template nucleic acid before and after the pause; identifying a first portion of the sequence read generated before the pause that is complementary to a second portion of the sequence read generated after the pause; and determining a likelihood that a hairpin loop formed by annealing of the first portion to the second portion was present in the template nucleic acid during the processing based at least upon the nucleotide composition of the first portion and the second portion.

[0021] In another aspect of the invention, methods for detecting binding of an agent to a single nucleic acid template are provided. In certain embodiments, such methods generally comprise providing the single nucleic acid template in complex with a polymerase; introducing a reaction mixture to the complex, wherein the reaction mixture comprises the agent; and monitoring synthesis of a polynucleotide by the polymerase, wherein the polynucleotide is complementary

to the single nucleic acid template, and wherein a change in the synthesis is indicative of binding of the agent to the single nucleic acid template. Examples of agents appropriate for use in such methods include, but are not limited to, transcription factors, polymerases, reverse transcriptases, histones, restriction enzymes, antibodies, nucleic acid binding proteins, and nucleic acid binding agents. Examples of single nucleic acid templates appropriate for use in such methods include, but are not limited to, double-stranded DNA, double-stranded RNA, single-stranded DNA, single-stranded RNA, DNA/RNA hybrids, and templates comprising both double-stranded and single-stranded regions.

[0022] In certain aspects of the invention, a consensus binding site of the agent is determined. This determination can comprise, e.g., performing a plurality of sequencing-by-synthesis reactions on a set of single nucleic acid templates in the presence of the agent to generate a set of binding-affected nascent polynucleotide sequences; performing a plurality of sequencing-by-synthesis reactions on the set of single nucleic acid templates in the absence of the agent to generate a set of full-length nascent polynucleotide sequences; analyzing the binding-affected nascent polynucleotide sequences to determine a location at which the agent bound the single nucleic acid template during the sequencing-by-synthesis reactions in the presence of the agent; and identifying a sequence common to the full-length nascent polynucleotide sequences at the location, thereby identifying the consensus binding site of the agent. In certain embodiments, the binding-affected nascent polynucleotide sequences are truncated nascent polynucleotide sequences; and in other embodiments, the binding-affected nascent polynucleotide sequences are nascent polynucleotide sequences whose synthesis was paused at the location at which the agent bound.

[0023] In yet further aspects of the invention, methods for detecting modifications in a single nucleic acid template during a sequencing-by-synthesis reaction are provided. For example, such a method can comprise providing the single nucleic acid template in complex with a polymerase; introducing a reaction mixture to the complex, wherein the reaction mixture comprises an agent that specifically binds to the modification; and monitoring synthesis of a polynucleotide by the polymerase, wherein the polynucleotide is complementary to the single nucleic acid template, and wherein a pause or cessation of the synthesis of the polynucleotide is indicative of binding of the agent to the single nucleic acid template, thereby detecting the modification in the single nucleic acid template. In certain embodiments, the modification is an 8-oxoG lesion and/or the agent is a protein is selected from the group consisting of hOGG1, FPG, yOGG1, AlkA, Nth, Nei, MutY, UDG, SMUG, TDG, NEIL, an antibody against 8-oxoG, or a binding domain thereof. In other embodiments, the modification is a methylated base and/or the agent is a protein selected from the group consisting of MECP2, MBD1, MBD2, MBD4,

UHRF1, an antibody against the methylated base, or a binding domain thereof. In further embodiments, the modification is a secondary structure formation in the nucleic acid template. Preferably, the complex is immobilized in an optical confinement. The template can comprise, e.g., single-stranded linear nucleic acid, single-stranded circular nucleic acid, double-stranded linear nucleic acid, double-stranded circular nucleic acid, or a combination thereof.

[0024] In certain embodiments, a modification in a template nucleic acid can be repaired by including components of damage repair machinery in the reaction mixture, e.g., during a sequencing-by-synthesis reaction. In certain embodiments, the readlength of the sequencing-by-synthesis reaction is longer than that for a further sequencing-by-synthesis reaction performed with the single nucleic acid template in complex with the polymerase, but absent the agent and the damage repair machinery.

[0025] In other aspects of the invention, methods for bypassing one or more modifications in a single nucleic acid template during a sequencing-by-synthesis reaction are provided. Certain exemplary methods include providing the single nucleic acid template in complex with a sequencing engine; introducing a reaction mixture to the complex, wherein the reaction mixture comprises a bypass polymerase; initiating the sequencing-by-synthesis reaction; monitoring synthesis of a polynucleotide by the sequencing engine, wherein the polynucleotide is complementary to the single nucleic acid template, and wherein a pause or cessation of the synthesis of the polynucleotide is indicative that the sequencing engine has encountered a modification in the single nucleic acid template; subsequently monitoring synthesis of the polynucleotide by the bypass polymerase, which is indicative that the modification is being bypassed; and repeating the monitoring steps each time a further modification is encountered in the single nucleic acid template, thereby bypassing one or more modifications in a single nucleic acid template during a sequencing-by-synthesis reaction. In certain embodiments, the bypass polymerase comprises a detectable label and detection of a signal from the detectable label during the sequencing-by-synthesis reaction is indicative that the bypass polymerase is actively synthesizing the polynucleotide. In preferred embodiments, the readlength of the sequencing-by-synthesis reaction is longer than that for a further sequencing-by-synthesis reaction performed with the single nucleic acid template in complex with the sequencing engine, but absent the bypass polymerase. In specific embodiments, the reaction mixture comprises multiple different bypass polymerases and a processivity factor. Preferably, at least one of the single nucleic acid template, the sequencing engine, and the bypass polymerase is immobilized, directly or indirectly, in an optical confinement. For example, the template can be immobilized by hybridization to an oligonucleotide primer immobilized in the optical confinement. In certain preferred embodiments, the single nucleic acid template is processed by the sequencing engine

multiple times at a single reaction site, and further wherein redundant sequence data is generated. Nucleic acid templates for use with the methods can be circular and/or can comprise multiple copies of a nucleic acid segment of interest. Further, in certain embodiments the sequencing-by-synthesis reaction generates a polynucleotide comprising multiple copies of a segment complementary to the segment of interest, and further wherein redundant sequence data is generated.

[0026] In further aspects, novel compositions are provided. For example, in certain embodiments a composition of the invention comprises a substrate having a reaction site that is optically resolvable from any other reaction site on the substrate; a single complex of a template and sequencing engine immobilized at the reaction site; a mixture of incorporatable nucleotides or nucleotide analogs; and at least one modification in the template nucleic acid, wherein the template at or proximal to the modification is processed differently than the template distal from the modification. In some embodiments the modification is a non-natural base in the template. The modification can be located in either a strand of the template nucleic acid that is complementary to a nascent strand synthesized by the sequencing engine, or a strand of the template nucleic acid that is displaced by the sequencing engine. In certain preferred embodiments, the template nucleic acid comprises internally complementary regions, and optionally, the modification is located within one of the internally complementary regions. Certain embodiments further comprise at least one type of non-incorporatable nucleotide analog. Certain embodiments comprise at least one type of non-natural incorporatable nucleotide analog. Preferably, one or more or all of the nucleotides or nucleotide analogs in a composition of the invention are tagged with distinct labels that distinguish different types of nucleotides or nucleotide analogs from one another. Compositions of the invention can also include an agent other than the sequencing engine that binds to the modification and/or chemically or enzymatically alters the modification. Preferably, compositions of the invention comprise a nascent strand generated by the sequencing engine, wherein the nascent strand is complementary to the template nucleic acid, and, optionally, comprises multiple copies of regions complementary to the template nucleic acid. Further, certain compositions comprise a nanometer-scale aperture in the substrate, where the reaction site is disposed within the nanometer-scale aperture, e.g., a zero-mode waveguide.

[0027] In further aspects of the invention, systems for identification of modifications within a nucleic acid template are provided. In certain preferred embodiments, a system of the invention comprises a solid support having a polymerase complex disposed thereon (e.g., at a reaction site, e.g., in a nanoscale aperture, e.g., in a zero-mode waveguide), the polymerase complex comprising a nucleic acid template comprising a modification; a mounting stage configured to

receive the solid support; an optical train positioned to be in optical communication with at least a portion of the solid support to detect signals emanating therefrom; a translation system operably coupled to the mounting stage or the optical train for moving one of the optical train and the solid support relative to the other; and a data processing system operably coupled to the optical train. Preferably, the polymerase complex comprises a polymerase enzyme that is actively processing the nucleic acid template. More preferably, the polymerase complex comprises a polymerase enzyme that is processively synthesizing a nascent strand by template-directed synthesis. In preferred embodiments, the optical train detects signals emanating from the solid support during the processing of the nucleic acid template.

[0028] In yet further aspects, the invention provides machine-implemented methods for transforming reaction data into modification detection data, wherein the reaction data is representative of a series of events during a sequencing-by-synthesis reaction wherein a nascent strand is synthesized based upon a nucleotide sequence of a template nucleic acid, and the modification detection data is representative of a presence of one or more modifications within a template nucleic acid. Preferably, one or more steps of the machine-implemented method are performed via a user interface implemented in a machine that comprises instructions stored in machine-readable medium and a processor that executes the instructions. In a final aspect of the invention, a computer program products are provided. In certain embodiments, machine-implemented methods for transforming reaction data comprise a classifier to distinguish between true incorporations and stochastic pulses, a segmenting algorithm based on a hidden Markov model architecture, and/or a segmenting algorithm based on a conditional random field framework. In certain specific embodiments, the methods identify regions in the template having a higher density of stochastic pulses than true incorporations. In certain specific embodiments, the methods identify regions in the template having higher IPD. Exemplary computer program products of the invention typically comprise a computer usable medium having a computer readable program code embodied therein, said computer readable program code adapted to be executed to implement the machine-implemented methods of the invention; and the machine-readable medium on which the results of one or more steps of the machine-implemented method are stored.

BRIEF DESCRIPTION OF THE DRAWINGS

[0029] Figure 1 provides an exemplary illustration of single-molecule, real-time (SMRT™) nucleic acid sequencing.

[0030] Figure 2 provides illustrative examples of various types of reaction data in the context of a pulse trace.

[0031] Figure 3 schematically illustrates a structural model of 5-MeC positioned one base in the 5' direction relative to the DNA polymerase active site.

[0032] Figure 4 illustrates an exemplary embodiment of five-base DNA methylation sequencing. Figure 4A depicts fragmentation of genomic DNA to generate the DNA template. Figure 4B illustrates DNA glycosylase excising a 5-MeC from the template.

[0033] Figure 5 provides an illustrative embodiment of a reaction comprising a linear template and a damage-binding agent that recognizes a lesion in a single-stranded template.

[0034] Figure 6 illustrates an embodiment of the invention comprising a circular template and a damage-binding agent that recognizes a lesion in a double-stranded template.

[0035] Figure 7 illustrates an observation of true incorporations (solid line) versus stochastic pulses (dashed line) across time.

[0036] Figure 8 provides an illustrative example of a simple hidden Markov model for classifying pause (P) versus sequencing (S) states within a sequencing trace.

[0037] Figure 9 provides an illustrative example of a system of the invention.

[0038] Figure 10A provides a schematic for exemplary template nucleic acids of the invention. Figure 10B provides graphs plotting interpulse duration for template nucleic acids as depicted in 10A.

[0039] Figure 11 provides a graph plotting IPD ratio against template position for a template nucleic acid comprising 5-methylcytosine modifications.

[0040] Figure 12 provides a graph plotting IPD ratio against template position for a template nucleic acid comprising 5-methylcytosine modifications.

[0041] Figure 13A provides a schematic for exemplary template nucleic acids of the invention. Figure 13B provides graphs plotting interpulse duration for template nucleic acids as depicted in 13A. Figure 13C provides a ROC curve for the data provided in 13B.

[0042] Figure 14 provides a graph plotting IPD ratio against template position for a template nucleic acid comprising N6-methyladenosine modifications.

[0043] Figure 15 provides a graph plotting IPD ratio against template position for a template nucleic acid comprising 5-hydroxymethylcytosine modifications.

[0044] Figure 16 provides a graph plotting pulse width ratio against template position for a template nucleic acid comprising 5-hydroxymethylcytosine modifications.

[0045] Figure 17 provides a graph plotting IPD ratio against template position for a template nucleic acid comprising 8-oxoguanosine modifications.

[0046] Figure 18 provides a graph plotting pulse width ratio against template position for a template nucleic acid comprising 8-oxoguanosine modifications.

DETAILED DESCRIPTION OF THE INVENTION

I. General

[0047] The present invention is generally directed to methods, compositions, and systems for detecting modifications within nucleic acid sequences, and in particularly preferred aspects, methylated nucleotides within sequence templates through the use of single molecule nucleic acid analysis. The ability to detect modifications within nucleic acid sequences is useful for mapping such modifications in various types and/or sets of nucleic acid sequences, e.g., across a set of mRNA transcripts, across a chromosomal region of interest, or across an entire genome. The modifications so mapped can then be related to transcriptional activity, secondary structure of the nucleic acid, siRNA activity, mRNA translation dynamics, kinetics and/or affinities of DNA- and RNA-binding proteins, and other aspects of nucleic acid (e.g., DNA and/or RNA) metabolism.

[0048] Although certain embodiments of the invention are described in terms of detection of modified nucleotides or other modifications in a single-stranded DNA molecule (e.g., a single-stranded template DNA), various aspects of the invention are applicable to many different types of nucleic acids, including e.g., single- and double-stranded nucleic acids that may comprise DNA, RNA (e.g., mRNA, siRNA, microRNA, rRNA, tRNA, snRNA, etc.), RNA-DNA hybrids, PNA, LNA, morpholino, and other RNA and/or DNA mimetics and derivatives thereof, and combinations of any of the foregoing. Nucleic acids for use with the methods, compositions, and systems provided herein may consist entirely of native nucleotides, or may comprise non-natural bases/nucleotides (e.g., synthetic and/or engineered) that may be paired with native nucleotides or may be paired with the same or a different non-natural base/nucleotide. In certain preferred embodiments, the nucleic acid comprises a combination of single-stranded and double-stranded regions, e.g., such as the templates described in U.S.S.N. 12/383,855 and 12/413,258, both filed on March 27, 2009 and incorporated herein by reference in their entireties for all purposes. In particular, mRNA modifications are difficult to detect by technologies that require reverse transcriptase PCR amplification because such treatment does not maintain the modification in the amplicons. The present invention provides methods for analyzing modifications in RNA molecules that do not require such amplification.

[0049] Generally speaking, the methods of the invention involve monitoring of an analytical reaction to collect "reaction data," wherein the reaction data is indicative of the progress of the reaction. Reaction data included data collected directly from the reaction, as well as the results of various manipulations of that directly collected data, any or a combination of which can serve as a signal for the presence of a modification in the template nucleic acid. For example, certain

types of reaction data are collected in real time during the course of the reaction, such as metrics related to reaction kinetics, processivity, signal characteristics, and the like. Signal characteristics vary depending on the type of analytical reaction being monitored. For example, some reactions use detectable labels to tag one or more reaction components, and signal characteristics for a detectable label include, but are not limited to, the type of signal (e.g., wavelength, charge, etc.) and the shape of the signal (e.g., height, width, curve, etc.). Further, signal characteristics for multiple signals (e.g., temporally adjacent signals) can also be used, including, e.g., the distance between signals during a reaction, the number of extra signals (e.g., that do not correspond to the progress of the reaction), internal complementarity, and the local signal context (i.e., one or more signal that precede and/or follow a given signal). For example, template-directed sequencing reactions often combine signal data from multiple nucleotide incorporation events to generate a sequence read for a nascent strand synthesized, and this sequence read is used to derive, e.g., by complementarity, the sequence of the template strand. Other types of reaction data are generated from statistical analysis of real time reaction data, including, e.g., accuracy, precision, conformance, etc. In some embodiments, data from a source other than the reaction being monitored is also used. For example, a sequence read generated during a nucleic acid sequencing reaction can be compared to sequence reads generated in replicate experiments, or to known or derived reference sequences from the same or a related biological source. Alternatively or additionally, a portion of a template nucleic acid preparation can be amplified using unmodified nucleotides and subsequently sequenced to provide an experimental reference sequence to be compared to the sequence of the original template in the absence of amplification. Although certain specific embodiments of the use of particular types of reaction data to detect certain kinds of modifications are described at length herein, it is to be understood that the methods, compositions, and systems are not limited to these specific embodiments. Different types of reaction data can be combined to detected various kinds of modifications, and in certain embodiments more than one type of modification can be detected and identified during a single reaction on a single template. Such variations to the detailed embodiments of the invention will be clear to one of ordinary skill based upon the teachings provided herein.

[0050] In certain embodiments, redundant sequence information is generated and analyzed to detect one or more modifications in a template nucleic acid. Redundancy can be achieved in various ways, including carrying out multiple sequencing reactions using the same original template, e.g., in an array format, e.g., a ZMW array. In some embodiments in which a lesion is unlikely to occur in all the copies of a given template, reaction data (e.g., sequence reads, kinetics, signal characteristics, signal context, and/or results from further statistical analyses)

generated for the multiple reactions can be combined and subjected to statistical analysis to determine a consensus sequence for the template. In this way, the reaction data from a region in a first copy of the template can be supplemented and/or corrected with reaction data from the same region in a second copy of the template. Similarly, a template can be amplified (e.g., via rolling circle amplification) to generate a concatemer comprising multiple copies of the template, and the concatemer can be subjected to sequencing, thereby generating a sequencing read that is internally redundant. As such, the sequence data from a first segment of the concatemer (corresponding to a first region of the template) can be supplemented and/or corrected with sequence data from a second segment of the concatemer also corresponding to the first region of the template. Alternatively or additionally, a template can be subjected to repeated sequencing reactions to generate redundant sequence information that can be analyzed to more thoroughly characterize the modification(s) present in the template.

[0051] The term "modification" as used herein is intended to refer not only to a chemical modification of a nucleic acids, but also to a variation in nucleic acid conformation, interaction of an agent with a nucleic acid (e.g., bound to the nucleic acid), and other perturbations associated with the nucleic acid. As such, a location or position of a modification is a locus (e.g., a single nucleotide or multiple contiguous or noncontiguous nucleotides) at which such modification occurs within the nucleic acid. For a double-stranded template, such a modification may occur in the strand complementary to a nascent strand synthesized by a polymerase processing the template, or may occur in the displaced strand. Although certain specific embodiments of the invention are described in terms of 5-methylcytosine detection, detection of other types of modified nucleotides (e.g., N^6-methyladenosine, N^3-methyladenosine, N^7-methylguanosine, 5-hydroxymethylcytosine, other methylated nucleotides, pseudouridine, thiouridine, isoguanosine, isocytosine, dihydrouridine, queuosine, wyosine, inosine, triazole, diaminopurine, 8-oxoguanosine, and 2'-O-methyl derivatives of adenosine, cytidine, guanosine, and uridine) is also contemplated. These and other modifications are known to those of ordinary skill in the art and are further described, e.g., in Narayan P, et al. (1987) Mol Cell Biol 7(4):1572-5; Horowitz S, et al. (1984) Proc Natl Acad Sci U.S.A. 81(18):5667-71; "RNA's Outfits: The nucleic acid has dozens of chemical costumes," (2009) C&EN; 87(36):65-68; Kriaucionis, et al. (2009) Science 324 (5929): 929–30; and Tahiliani, et al. (2009) Science 324 (5929): 930–35; Matray, et al. (1999) Nature 399(6737):704-8; Ooi, et al. (2008) Cell 133: 1145-8; Petersson, et al. (2005) J Am Chem Soc. 127(5):1424-30; Johnson, et al. (2004) 32(6):1937-41; Kimoto, et al. (2007) Nucleic Acids Res. 35(16):5360-9; Ahle, et al. (2005) Nucleic Acids Res 33(10):3176; Krueger, et al., Curr Opinions in Chem Biology 2007, 11(6):588); Krueger, et al. (2009) Chemistry & Biology 16(3):242; McCullough, et al. (1999)

Annual Rev of Biochem 68:255; and Liu, et al. (2003) Science 302(5646):868-71, the disclosures of which are incorporated herein by reference in their entireties for all purposes. Modifications further include the presence of non-natural base pairs in the template nucleic acid, including but not limited to hydroxypyridone and pyridopurine homo- and hetero-base pairs, pyridine-2,6-dicarboxylate and pyridine metallo-base pairs, pyridine-2,6-dicarboxamide and a pyridine metallo-base pairs, metal-mediated pyrimidine base pairs T-Hg(II)-T and C-Ag(I)-C, and metallo-homo-basepairs of 2,6-bis(ethylthiomethyl)pyridine nucleobases Spy, and alkyne-, enamine-, alcohol-, imidazole-, guanidine-, and pyridyl-substitutions to the purine or pyridimine base (Wettig, et al. (2003) J Inorg Biochem 94:94-99; Clever, et al. (2005) Angew Chem Int Ed 117:7370-7374; Schlegel, et al. (2009) Org Biomol Chem 7(3):476-82; Zimmerman, et al. (2004) Bioorg Chem 32(1):13-25; Yanagida, et al. (2007) Nucleic Acids Symp Ser (Oxf) 51:179-80; Zimmerman (2002) J Am Chem Soc 124(46):13684-5; Buncel, et al. (1985) Inorg Biochem 25:61-73; Ono, et al. (2004) Angew Chem 43:4300-4302; Lee, et al. (1993) Biochem Cell Biol 71:162-168; Loakes, et al. (2009), Chem Commun 4619-4631; and Seo, et al. (2009) J Am Chem Soc 131:3246-3252, all incorporated herein by reference in their entireties for all purposes). Other types of modifications include, e.g, a nick, a missing base (e.g., apurinic or apyridinic sites), a pyrimidine dimer (e.g., thymine dimer or cyclobutane pyrimidine dimer), a cis-platin crosslinking, oxidation damage, hydrolysis damage, other methylated bases, bulky DNA base adducts, photochemistry reaction products, interstrand crosslinking products, mismatched bases, and other types of "damage" to the nucleic acid. As such, certain embodiments described herein refer to "damage" and such damage is also considered a modification of the nucleic acid in accordance with the present invention. Modified nucleotides can be caused by exposure of the DNA to radiation (e.g., UV), carcinogenic chemicals, crosslinking agents (e.g., formaldehyde), certain enzymes (e.g., nickases, glycosylases, exonucleases, methylases, other nucleases, etc.), viruses, toxins and other chemicals, thermal disruptions, and the like. *In vivo*, DNA damage is a major source of mutations leading to various diseases including cancer, cardiovascular disease, and nervous system diseases (see, e.g., Lindahl, T. (1993) Nature 362(6422): 709-15, which is incorporated herein by reference in its entirety for all purposes). The methods and systems provided herein can also be used to detect various conformations of DNA, in particular, secondary structure forms such as hairpin loops, stem-loops, internal loops, bulges, pseudoknots, base-triples, and the like; and are also useful for detection of agents interacting with the nucleic acid, e.g., bound proteins or other moieties.

[0052] In certain aspects, methods, compositions, and systems for detection and/or reversal of modifications in a template for single-molecule sequencing are provided, as well as determination of their location (i.e. "mapping") within a nucleic acid molecule. In certain

preferred embodiments, high-throughput, real-time, single-molecule, template-directed sequencing assays are used to detect the presence of such damage sites and to determine their location on the DNA template, e.g., by monitoring the progress and/or kinetics of a polymerase enzyme processing the template. For example, when a polymerase enzyme encounters certain types of damage or other modifications in a DNA template, the progress of the polymerase can be temporarily or permanently blocked, e.g., resulting in a paused or dissociated polymerase. As such, the detection of a pause in or termination of nascent strand synthesis is indicative of the presence of such damage or lesion. By analysis of the sequence reads produced prior to the pause or stop in synthesis, and alternatively or additionally after reinitiation of synthesis, one can map the site of the damage or lesion on the template. Since different types of lesions can have different effects on the progress of the polymerase on the substrate, in certain cases the behavior of the polymerase on the template not only informs as to where the lesion occurs, but also what type of lesion is present. Further, in certain embodiments a modification may be bypassed by incorporation of a non-nucleotide binding partner with the lesion in the template strand. For example, abasic sites (e.g., produced by glycosylases) can be "paired" with pyrenes or other similar analogs. (See, e.g., Matray, et al. (1999) Nature 399(6737): 704-8). Such an analog can also be labeled with a detectable label optically distinguishable from those on the nucleotides in the reaction mixture to allow optical detection of its incorporation. Certain aspects of the invention provide a means for reversing such modifications in real time, thereby allowing reinitation of the sequencing reaction and continued generation of sequence information for the template nucleic acid. Such methods can additionally be used to study effects of various agents (e.g., drugs, chemicals, enzymes, etc.) and reaction conditions on the creation and/or repair of such lesions and/or damage, as described elsewhere herein. These and other aspects of the invention are described in greater detail in the description and examples that follow.

II. Single Molecule Sequencing

[0053] In certain aspects of the invention, single molecule real time sequencing systems are applied to the detection of modified nucleic acid templates through analysis of the sequence and/or kinetic data derived from such systems. In particular, modifications in a template nucleic acid strand alter the enzymatic activity of a nucleic acid polymerase in various ways, e.g., by increasing the time for a bound nucleotide to be incorporated and/or increasing the time between incorporation events. In certain embodiments, polymerase activity is detected using a single molecule nucleic acid sequencing technology. In certain embodiments, polymerase activity is detected using a nucleic acid sequencing technology that detects incorporation of nucleotides into a nascent strand in real time. In preferred embodiments, a single molecule nucleic acid

sequencing technology is capable of real-time detection of nucleotide incorporation events. Such sequencing technologies are known in the art and include, e.g., the SMRT™ sequencing and nanopore sequencing technologies. For more information on nanopore sequencing, see, e.g., U.S. Patent No. 5,795,782; Kasianowicz, et al. (1996) Proc Natl Acad Sci USA 93(24):13770-3; Ashkenas, et al. (2005) Angew Chem Int Ed Engl 44(9):1401-4; Howorka, et al. (2001) Nat Biotechnology 19(7):636-9; and Astier, et al. (2006) J Am Chem Soc 128(5):1705-10, all of which are incorporated herein by reference in their entireties for all purposes. With regards to nucleic acid sequencing, the term "template" refers to a nucleic acid molecule subjected to template-directed synthesis of a nascent strand. A template may comprise DNA, RNA, or mimetics or derivatives thereof. Further, a template may be single-stranded, double-stranded, or may comprise both single- and double-stranded regions. A modification in a double-stranded template may be in the strand complementary to the newly synthesized nascent strand, or may by in the strand identical to the newly synthesized strand, i.e., the strand that is displaced by the polymerase.

[0054] The preferred direct methylation sequencing described herein may generally be carried out using single molecule real time sequencing systems, i.e., that illuminate and observe individual reaction complexes continuously over time, such as those developed for SMRT™ DNA sequencing (see, e.g., P. M. Lundquist, et al., *Optics Letters* **2008**, *33*, 1026, which is incorporated herein by reference in its entirety for all purposes). The foregoing SMRT™ sequencing instrument generally detects fluorescence signals from an array of thousands of ZMWs simultaneously, resulting in highly parallel operation. Each ZMW, separated from others by distances of a few micrometers, represents an isolated sequencing chamber.

[0055] Detection of single molecules or molecular complexes in real time, e.g., during the course of an analytical reaction, generally involves direct or indirect disposal of the analytical reaction such that each molecule or molecular complex to be detected is individually resolvable. In this way, each analytical reaction can be monitored individually, even where multiple such reactions are immobilized on a single substrate. Individually resolvable configurations of analytical reactions can be accomplished through a number of mechanisms, and typically involve immobilization of at least one component of a reaction at a reaction site. Various methods of providing such individually resolvable configurations are known in the art, e.g., see European Patent No.1105529 to Balasubramanian, et al.; and Published International Patent Application No. WO 2007/041394, the full disclosures of which are incorporated herein by reference in their entireties for all purposes. A reaction site on a substrate is generally a location on the substrate at which a single analytical reaction is performed and monitored, preferably in real time. A reaction site may be on a planar surface of the substrate, or may be in an aperture in

the surface of the substrate, e.g., a well, nanohole, or other aperture. In preferred embodiments, such apertures are "nanoholes," which are nanometer-scale holes or wells that provide structural confinement of analytic materials of interest within a nanometer-scale diameter, e.g., ~1-300 nm. In some embodiments, such apertures comprise optical confinement characteristics, such as zero-mode waveguides, which are also nanometer-scale apertures and are further described elsewhere herein. Typically, the observation volume (i.e., the volume within which detection of the reaction takes place) of such an aperture is at the attoliter (10^{-18} L) to zeptoliter (10^{-21} L) scale, a volume suitable for detection and analysis of single molecules and single molecular complexes.

[0056] The immobilization of a component of an analytical reaction can be engineered in various ways. For example, an enzyme (e.g., polymerase, reverse transcriptase, kinase, etc.) may be attached to the substrate at a reaction site, e.g., within an optical confinement or other nanometer-scale aperture. In other embodiments, a substrate in an analytical reaction (for example, a nucleic acid template, e.g., DNA, RNA, or hybrids, analogs, and mimetics thereof, or a target molecule for a kinase) may be attached to the substrate at a reaction site. Certain embodiments of template immobilization are provided, e.g., in U.S. Patent Application No. 12/562,690, filed September 18, 2009 and incorporated herein by reference in its entirety for all purposes. One skilled in the art will appreciate that there are many ways of immobilizing nucleic acids and proteins into an optical confinement, whether covalently or non-covalently, via a linker moiety, or tethering them to an immobilized moiety. These methods are well known in the field of solid phase synthesis and micro-arrays (Beier et al., Nucleic Acids Res. 27:1970-1-977 (1999)). Non-limiting exemplary binding moieties for attaching either nucleic acids or polymerases to a solid support include streptavidin or avidin/biotin linkages, carbamate linkages, ester linkages, amide, thiolester, (N)-functionalized thiourea, functionalized maleimide, amino, disulfide, amide, hydrazone linkages, among others. Antibodies that specifically bind to one or more reaction components can also be employed as the binding moieties. In addition, a silyl moiety can be attached to a nucleic acid directly to a substrate such as glass using methods known in the art.

[0057] In some embodiments, a nucleic acid template is immobilized onto a reaction site (e.g., within an optical confinement) by attaching a primer comprising a complementary region at the reaction site that is capable of hybridizing with the template, thereby immobilizing it in a position suitable for monitoring. In certain embodiments, an enzyme complex is assembled in an optical confinement, e.g., by first immobilizing an enzyme component. In other embodiments, an enzyme complex is assembled in solution prior to immobilization. Where desired, an enzyme or other protein reaction component to be immobilized may be modified to contain one or more

epitopes for which specific antibodies are commercially available. In addition, proteins can be modified to contain heterologous domains such as glutathione S-transferase (GST), maltose-binding protein (MBP), specific binding peptide regions (see e.g., U.S. Patent Nos. 5,723,584, 5,874,239 and 5,932,433), or the Fc portion of an immunoglobulin. The respective binding agents for these domains, namely glutathione, maltose, and antibodies directed to the Fc portion of an immunoglobulin, are available and can be used to coat the surface of an optical confinement of the present invention.The binding moieties or agents of the reaction components they immobilize can be applied to a support by conventional chemical techniques which are well known in the art. In general, these procedures can involve standard chemical surface modifications of a support, incubation of the support at different temperature levels in different media comprising the binding moieties or agents, and possible subsequent steps of washing and cleaning.

[0058] In some embodiments, a substrate comprising an array of reaction sites is used to monitor multiple biological reactions, each taking place at a single one of the reaction sites. Various means of loading multiple biological reactions onto an arrayed substrate are known to those of ordinary skill in the art and are described further, e.g., in USSN 61/072,641, incorporated herein by reference in its entirety for all purposes. For example, basic approaches include: creating a single binding site for a reaction component at the reaction site; removing excess binding sites at the reaction site via catalytic or secondary binding methods; adjusting the size or charge of the reaction component to be immobilized; packaging or binding the reaction component within (or on) a particle (e.g., within a viral capsid), where a single such particle fits into the relevant reaction site (due to size or charge of the particle and/or observation volume); using non-diffusion limited loading; controllably loading the reaction component (e.g., using microfluidic or optical or electrical control); sizing or selecting charges in the reaction sites/observation volumes (e.g., the sizes of optical confinements in an array) to control which reaction components will fit (spatially or electrostatically) into which reaction sites/observation volumes; iterative loading of reaction components, e.g., by masking active sites between loading cycles; enriching the activity of the reaction components that are loaded; using self-assembling nucleic acids to sterically control loading; adjusting the size of the reaction site/observation volume; and many others. Such methods and compositions provide for the possibility of completely loading single-molecule array reaction sites (instead of about 30% of such sites as occurs in "Poisson limited" loading methods) with single reaction components (e.g., molecular complexes).

[0059] In preferred aspects, the methods, compositions, and systems provided herein utilize optical confinements to facilitate single molecule resolution of analytical reactions. In preferred

embodiments, such optical confinements are configured to provide tight optical confinement so only a small volume of the reaction mixture is observable. Some such optical confinements and methods of manufacture and use thereof are described at length in, e.g., U.S. Patent. Nos. 7,302,146. 7,476,503, 7,313,308, 7,315,019, 7,170,050, 6,917,726, 7,013,054, 7,181,122, and 7,292,742; U.S. Patent Publication Nos. 20080128627, 20080152281, and 200801552280; and U.S.S.N. 11/981,740 and 12/560,308, all of which are incorporated herein by reference in their entireties for all purposes.

[0060] Where reaction sites are located in optical confinements, the optical confinements can be further tailored in various ways for optimal confinement of an analytical reaction of interest. In particular, the size, shape, and composition of the optical confinement can be specifically designed for containment of a given enzyme complex and for the particular label and illumination scheme used.

[0061] In certain preferred embodiments of the invention, single-molecule real-time sequencing systems already developed are applied to the detection of modified nucleic acid templates through analysis of the sequence and kinetic data derived from such systems. As described below, methylated cytosine and other modifications in a template nucleic acid will alter the enzymatic activity of a polymerase processing the template nucleic acid. In certain embodiments, polymerase kinetics in addition to sequence read data are detected using a single molecule nucleic acid sequencing technology, e.g., the SMRT™ sequencing technology developed by Pacific Biosciences (Eid, J. et al. (2009) *Science* **2009**, *323*, 133, the disclosure of which is incorporated herein by reference in its entirety for all purposes). This technique is capable of long sequencing reads and provides high-throughput methylation profiling even in highly repetitive genomic regions, facilitating *de novo* sequencing of modifications such as methylated bases. SMRT™ sequencing systems typically utilize state-of-the-art single-molecule detection instruments, production-line nanofabrication chip manufacturing, organic chemistry, protein mutagenesis, selection and production facilities, and software and data analysis infrastructures.

[0062] Certain preferred methods of the invention employ real-time sequencing of single DNA molecules (Eid, et al., *supr*a), with intrinsic sequencing rates of several bases per second and average read lengths in the kilobase range. In such sequencing, sequential base additions catalyzed by DNA polymerase into the growing complementary nucleic acid strand are detected with fluorescently labeled nucleotides. The kinetics of base additions and polymerase translocation are sensitive to the structure of the DNA double-helix, which is impacted by the presence of base modifications, e.g, 5-MeC and other perturbations (secondary structure, bound agents, etc.) in the template. By monitoring the activity of DNA polymerase during sequencing,

sequence read information and base modifications can be simultaneously detected. Long, continuous sequence reads that are readily achievable using SMRT™ sequencing facilitate modification (e.g., methylation) profiling in low complexity regions that are inaccessible to some technologies, such as certain short-read sequencing technologies. Carried out in a highly parallel manner, methylomes can be sequenced directly, with single base-pair resolution and high throughput.

[0063] The principle of SMRT™ sequencing is illustrated in Figure 1. Two important technology components of certain embodiments of this process are: (i) optical confinement technology that allows single-molecule detection at concentrations of labeled nucleotides relevant to the enzyme, and (ii) phospholinked nucleotides that enable observation of uninterrupted polymerization.

[0064] In preferred embodiments, optical confinements are ZMW nanostructures, preferably in an arrayed format. Typically, ZMWs arrays comprise dense arrays of holes, ~100 nm in diameter, fabricated in a ~100 nm thick metal film deposited on a transparent substrate (e.g., silicon dioxide). These structures are further described in the art, e.g., in M. J. Levene, et al., *Science* **2003**, *299*, 682 ; and M. Foquet, et al., *J. Appl. Phys.* **2008**, *103*, 034301, the disclosures of which are incorporated herein by reference in their entireties for all purposes. Each ZMW becomes a nanophotonic visualization chamber for recording an individual polymerization reaction, providing a detection volume of just 100 zeptoliters (10^{-21} liters). This volume represents a ~1000-fold improvement over diffraction-limited confocal microscopy, facilitating observation of single incorporation events against the background created by the relatively high concentration of fluorescently labeled nucleotides. Polyphosphonate and silane-based surface coatings mediate enzyme immobilization to the transparent floor of the ZMW while blocking non-specific attachments to the metal top and side wall surfaces (Eid, et al., *supra*; and J. Korlach, et al., *Proc Natl Acad Sci U S A* **2008**, *105*, 1176, the disclosures of which are incorporated herein by reference in their entireties for all purposes). While certain methods described herein involve the use of ZMW confinements, it will be readily understood by those of ordinary skill in the art upon review of the teachings herein that these methods may also be practiced using other reaction formats, e.g., on planar substrates or in nanometer-scale apertures other than zero-mode waveguides. (See, e.g., U.S.S.N. 12/560,308, filed September 15, 2009; and U.S. Patent Publication No. 20080128627, incorporated herein *supra*.)

[0065] The second important component is phospholinked nucleotides for which a detectable label (e.g., comprising a fluorescent dye) is attached to the terminal phosphate rather than the base (Figure 1). (See, e.g., J. Korlach, et al., *Nucleos. Nucleot. Nucleic Acids* **2008**, *27*, 1072, which is incorporated herein by reference in its entirety for all purposes.) 100% replacement of

unmodified nucleotides by phospholinked nucleotides is performed, and the enzyme cleaves away the label as part of the incorporation process, leaving behind a completely natural, double-stranded nucleic acid product. Each of the four different nucleobases is labeled with a distinct detectable label to discriminate base identities during incorporation events, thus enabling sequence determination of the complementary DNA template. During incorporation, the enzyme holds the labeled nucleotide in the ZMW's detection volume for tens of milliseconds, orders of magnitude longer than the average diffusing nucleotide is present. Signal (e.g., fluorescence) is emitted continuously from the detectable label during the duration of incorporation, causing a detectable pulse of increased fluorescence in the corresponding color channel. The pulse is terminated naturally by the polymerase releasing the pyrophosphate-linker-label group. The polymerase then translocates to the next base, and the process repeats. As shown in Figure 1A, single DNA polymerase molecules with bound DNA template are attached to a substrate, e.g., at the bottom of each zero-mode waveguide. Polymerization of the complementary DNA strand is observed in real time by detecting fluorescently labeled nucleotides. Reactions steps involved in SMRT™ sequencing are as follows: Step 1: The DNA template/primer/polymerase complex is surrounded by diffusing fluorescently labeled nucleotides which probe the active site. Step 2: A labeled nucleotide makes a cognate binding interaction with the next base in the DNA template that lasts for tens of milliseconds, during which fluorescence is emitted continuously. Step 3: The polymerase incorporates the nucleotide into the growing nucleic acid chain, thereby cleaving the α-β phosphodiester bond, followed by release of the nucleotide. Steps 4-5: The process repeats. A prophetic trace is shown in Figure 1B that comprises each step shown in 1A. At steps 2 and 4, a fluorescent signal is emitted during binding and incorporation of a nucleotide into the growing nucleic acid chain, and monitoring of these fluorescent signals provides a sequence of nucleotide incorporations that can be used to derive the sequence of the template nucleic acid. For example, a 5'-G-A-3' sequence in the growing chain indicates a 5'-T-C-3' sequence in the complementary template strand.

[0066] As described above, reaction data is indicative of the progress of a reaction and can serve as a signal for the presence of a modification in the template nucleic acid. Reaction data in single molecule sequencing reaction reactions using fluorescently labeled bases is generally centered around characterization of detected fluorescence pulses, a series of successive pulses ("pulse trace" or one or more portions thereof), and other downstream statistical analyses of the pulse and trace data. Fluorescence pulses are characterized not only by their spectrum, but also by other metrics including their duration, shape, intensity, and by the interval between successive pulses (see, e.g., Eid, et al., *supra*; and U.S. Patent Publication No. 20090024331, incorporated herein by reference in its entirety for all purposes). While not all of these metrics are generally

required for sequence determination, they add valuable information about the processing of a template, e.g., the kinetics of nucleotide incorporation and DNA polymerase processivity and other aspects of the reaction. Further, the context in which a pulse is detected (i.e., the one or more pulses that precede and/or follow the pulse) can contribute to the identification of the pulse. For example, the presence of certain modifications alters not only the processing of the template at the site of the modification, but also the processing of the template upstream and/or downstream of the modification. For example, the presence of modified bases in a template nucleic acid has been shown to change the width of a pulse and/or the interpulse duration (IPD), either at the modified base or at one or more positions proximal to it. A change in pulse width may or may not be accompanied by a change in IPD. Figure 2 provides illustrative examples of various types of reaction data in the context of a pulse trace including IPD, pulse width (PW), pulse height (PH), and context. Figure 2A illustrates these reaction data on a pulse trace generated on an unmodified template, and Figure 2B illustrates how the presence of a modification (5-MeC) can elicit a change in one of these reaction data (IPD) to generate a signal (increased IPD) indicative of the presence of the modification.

[0067] In yet further embodiments, reaction data is generated by analysis of the pulse and trace data to determine error metrics for the reaction. Such error metrics include not only raw error rate, but also more specific error metrics, e.g., identification of pulses that did not correspond to an incorporation event, incorporations that were not accompanied by a detected pulse, incorrect incorporation events, and the like. Any of these error metrics, or combinations thereof, can serve as a signal indicative of the presence of one or more modifications in the template nucleic acid. In some embodiments, such analysis involves comparison to a reference sequence and/or comparison to replicate sequence information from the same or an identical template, e.g., using a standard or modified multiple sequence alignment. Certain types of modifications cause an increase in one or more error metrics. For example, some modifications can be "paired" with more than one type of incoming nucleotide or analog thereof, so replicate sequence reads for the region comprising the modification will show variable base incorporation opposite such a modification. Such variable incorporation is thereby indicative of the presence of the modification. Certain types of modifications cause an increase in one or more error metrics proximal to the modification, e.g., immediately upstream or downstream. The error metrics at a locus or within a region of a template are generally indicative of the type of modification(s) present at that locus or in that region of the template, and therefore serve as a signal of such modification(s). In preferred embodiments, at least some reaction data is collected in real time during the course of the reaction, e.g., pulse and/or trace characteristics.

[0068] Although described herein primarily with regards to fluorescently labeled nucleotides, other types of detectable labels and labeling systems can also be used with the methods, compositions, and systems described herein including, e.g., quantum dots, surface enhanced Raman scattering particles, scattering metallic nanoparticles, FRET systems, intrinsic fluorescence, non-fluorescent chromophores, and the like. Such labels are generally known in the art and are further described in Provisional U.S. Patent Application No. 61/186,661, filed June 12, 2009; U.S. Patent Nos. 6,399,335, 5,866,366, 7,476,503, and 4,981,977; U.S. Patent Pub. No. 2003/0124576; U.S.S.N. 61/164567; WO 01/16375; Mujumdar, et al Bioconjugate Chem. 4(2):105-111, 1993; Ernst, et al, Cytometry 10:3-10, 1989; Mujumdar, et al, Cytometry 10:1119, 1989; Southwick, et al, Cytometry 11:418-430, 1990; Hung, et al, Anal. Biochem. 243(1):15-27, 1996; Nucleic Acids Res. 20(11):2803-2812, 1992; and Mujumdar, et al, Bioconjugate Chem. 7:356-362, 1996; Intrinsic Fluorescence of Proteins, vol. 6, publisher: Springer US, ©2001; Kronman, M.J. and Holmes, L.G. (2008) Photochem and Photobio 14(2): 113-134; Yanushevich, Y.G., et al. (2003) Russian J. Bioorganic Chem 29(4) 325-329; and Ray, K., et al. (2008) J. Phys. Chem. C 112(46): 17957-17963, all of which are incorporated herein by reference in their entireties for all purposes. Many such labeling groups are commercially available, e.g., from the Amersham Biosciences division of GE Healthcare, and Molecular Probes/Invitrogen Inc. (Carlsbad, CA)., and are described in 'The Handbook — A Guide to Fluorescent Probes and Labeling Technologies, Tenth Edition' (2005) (available from Invitrogen, Inc./Molecular Probes and incorporated herein in its entirety for all purposes). Further, a combination of the labeling strategies described herein and known in the art for labeling reaction components can be used.

[0069] Various strategies, methods, compositions, and systems are provided herein for detecting modifications in a nucleic acid, e.g., during real-time nascent strand synthesis. For example, since DNA polymerases can typically bypass 5-MeC in a template nucleic acid and properly incorporate a guanine in the complementary strand opposite the 5-MeC, additional strategies are desired to detect such altered nucleotides in the template. Various such strategies are provided herein, such as, e.g., a) modification of the polymerase to introduce an specific interaction with the modified nucleotide; b) detecting variations in enzyme kinetics, e.g., pausing; c) use of a detectable and optionally modified nucleotide analog that specifically base-pairs with the modification and is potentially incorporated into the nascent strand; d) chemical treatment of the template prior to sequencing that specifically alters 5-MeC sites in the template; e) use of a protein that specifically binds to the modification in the template nucleic acid, e.g., delaying or blocking progression of a polymerase during replication; and f) use of sequence context (e.g., the higher frequency of 5-MeC nucleotides in CpG islands) to focus modification

detection efforts on regions of the template that are more likely to contain such a modification (e.g., GC-rich regions for 5-MeC detection). These strategies may be used alone or in combination to detect 5-MeC sites in a template nucleic acid during nascent strand synthesis.

III. Polymerase Modifications

[0070] Various different polymerases may be used in template-directed sequence reactions, e.g., those described at length, e.g., in U.S. Pat. No. 7,476,503, the disclosure of which is incorporated herein by reference in its entirety for all purposes. In brief, the polymerase enzymes suitable for the present invention can be any nucleic acid polymerases that are capable of catalyzing template-directed polymerization with reasonable synthesis fidelity. The polymerases can be DNA polymerases or RNA polymerases (including, e.g., reverse transcriptases), a thermostable polymerase or a thermally degradable polymerase, wildtype or modified. In some embodiments, the polymerases exhibit enhanced efficiency as compared to the wildtype enzymes for incorporating unconventional or modified nucleotides, *e.g.*, nucleotides linked with fluorophores. In certain preferred embodiments, the methods are carried out with polymerases exhibiting a high degree of processivity, *i.e.*, the ability to synthesize long stretches (e.g., over about 10 kilobases) of nucleic acid by maintaining a stable nucleic acid/enzyme complex. In certain preferred embodiments, sequencing is performed with polymerases capable of rolling circle replication. A preferred rolling circle polymerase exhibits strand-displacement activity, and as such, a single circular template can be sequenced repeatedly to produce a sequence read comprising multiple copies of the complement of the template strand by displacing the nascent strand ahead of the translocating polymerase. Since the methods of the invention can increase processivity of the polymerase by removing lesions that block continued polymerization, they are particularly useful for applications in which a long nascent strand is desired, e.g. as in the case of rolling-circle replication. Non-limiting examples of rolling circle polymerases suitable for the present invention include but are not limited to T5 DNA polymerase, T4 DNA polymerase holoenzyme, phage M2 DNA polymerase, phage PRD1 DNA polymerase, Klenow fragment of DNA polymerase, and certain polymerases that are modified or unmodified and chosen or derived from the phages Φ29 (Phi29), PRD1, Cp-1, Cp-5, Cp-7, Φ15, Φ1, Φ21, Φ25, BS 32 L17, PZE, PZA, Nf, M2Y (or M2), PR4, PR5, PR722, B103, SF5, GA-1, and related members of the *Podoviridae* family. In certain preferred embodiments, the polymerase is a modified Phi29 DNA polymerase, e.g., as described in U.S. Patent Publication No. 20080108082, incorporated herein by reference in its entirety for all purposes. Additional polymerases are provided, e.g., in U.S.S.N. 11/645,125, filed December 21, 2006; 11/645,135, filed December 21, 2006; 12/384,112, filed March 30, 2009; and 61/094,843, filed September 5,

2008; as well as in U.S. Patent Publication No. 20070196846, the disclosures of which are incorporated herein by reference in their entireties for all purposes.

[0071] Further optimization is achieved through improvement of enzyme kinetics, either through the screening of polymerase libraries and/or the engineering of polymerases, which include, e.g., DNA polymerases, RNA polymerases, reverse transcriptases, and the like. In particular, DNA polymerases may be screened to identify those that have desirable properties for detection of nucleic acid modifications described herein. Further, polymerases can be engineered through directed mutagenesis of one or more residues involved in various aspects of template-directed nascent strand synthesis. For example, careful examination of the crystal structure of the polymerase-DNA-nucleotide complex for certain polymerase enzymes has shown that the polymerase rotates and flips out the base on the single-stranded region of the template DNA that is adjacent to the active site in the 5' direction (i.e., the base in the "-1" position). During the subsequent DNA translocation process, this base is flipped into the active site. As such, amino acids that interact with a modified base in the -1 position or during the subsequent tranlocation can be altered or substituted to increase the enzyme's sensitivity to the modified base. In fact, any protein residues that come into close contact with a modification in the template are candidates for substitution or alteration. For example, family B polymerases mostly contain replicative polymerases and include the major eukaryotic DNA polymerases α, δ, ε, and also DNA polymerase ζ. Family B also includes DNA polymerases encoded by some bacteria and bacteriophages, e.g., T4, Phi29, and RB69 bacteriophages. Most family B polymerases share common structural features for DNA binding, and the residues along the DNA primer-template junction and the residues around the base binding pocket at the -1 location (pre-insertion position) can be mutated and the resulting mutants screened for enhanced response to a modification of interest. Specifically, when 5-MeC is in the "flipped-out" (-1) position, it is surrounded by several Φ29 polymerase amino acid residues, such as K392 and K422 which are positioned close to the methyl group (Figure 3). Mutations such as K392R/W/M and K422R/W/M that substitute the native lysine residue with amino acids with larger side chains (e.g., arginine, tryptophan, or methionine) may increase the polymerase's sensitivity to modified bases, potentially delaying the translocation step and slowing incorporation of the complementary dGTP. This is schematically illustrated with reference to Figure 3, which shows a structural model of 5-MeC positioned one base in the 5' direction relative to the Φ29 DNA polymerase active site. As shown, two polymerase residues, K392 and K422, are close to the methyl group of 5-MeC. These two residues are potential targets for site-specific mutagenesis to amino acids with larger side chains that will interact sterically with the methyl group. 5-MeC is

shown as the chemical structure in the right center, while the K392 and K422 residues are left center.

[0072] Any residues that come in close contact with a modified base are candidates for mutation. For example, using molecular morphing and energy minimization to model the translocation path of 5-MeC, a number of protein residues in the Φ29 polymerase have been identified within 5 Å of the methyl moiety as it is being flipped into the active site during the translocation step. The following groups of residues (listed in the order of the translocation path) are targets of mutations to residues with larger side chains: I93, M188, K392, V399, T421, K422; S95, Y101, M102; Q99, L123, K124, T189, A190; G191, S388; P127, L384, N387, S388; and L389, Y390, G391. In particular, I93Y and V399Y may introduce a 5-methylcytosine specific binding region, analogous to those shown by the crystal structures of the SRA/5-methylcytosine binding complex. For example, see G. V. Avvakumov, et al., *Nature* **2008**, *455*, 822; and H. Hashimoto, et al., *Nature* **2008**, *455*, 826, the disclosures of both of which are incorporated herein by reference in their entireties for all purposes. Although the residues identified above are specific to the Φ29 polymerase, one of ordinary skill will readily recognize that the structural similarity between the family B polymerases, and to a lesser extent family A polymerases and other polymerases, allows identification of homologous positions on related polymerases as targets for mutation based on the teachings herein.

[0073] In addition to the foregoing, additional improvements are derived from a molecular evolution program using these polymerases to enhance their ability to sense 5-MeC. Such programs have already been used to successfully improve large numbers of different enzymes for a variety of applications, including improving DNA polymerases for sequencing. Such methods may include diversification of the amino acid sequence space by mutagenic PCR and DNA shuffling, and/or yeast displays for expression and selection (see, e.g., S. A. Gai, et al., *Curr Opin Struct Biol* **2007**, *17*, 467; and D. Lipovsek, et al., *Chem Biol* **2007**, *14*, 1176, which are incorporated herein by reference in their entireties for all purposes, in which ~10^4 copies of a recombinant protein are displayed on the surface of a single yeast cell carrying the transgene for the protein. The genotype-phenotype linkage is provided by the yeast cell, but no protein purification is necessary as the displayed proteins have the same properties as bulk solutions of polymerase. With the available infrastructure in house, this program can be initiated without startup costs as soon as polymerase candidates emerge.

IV. Secondary Structure Detection

[0074] During single molecule sequencing as described *supra*, an otherwise highly processive trace is sometimes interrupted by a long pause. Such pausing can be caused by secondary structure, e.g., a hairpin loop, in the template strand. In certain aspects, the invention provides methods for not only identifying secondary structure in a template nucleic acid, but also for improving the overall accuracy of single molecule sequencing.

[0075] In certain embodiments, a sequencing read or "trace" is generated by subjecting a template nucleic acid to a real-time, template-directed sequencing reaction. The trace is examined to identify long pauses by finding portions of the trace at which the interpulse duration (IPD) is significantly longer than the average IPD. For example, find pauses that are at least 2-, 3-, 5-, 10-, or 20-times longer than the average IPD. In some embodiments, an IPD averaged over a few neighboring bases is used, in other embodiments an IPD averaged over a window of about 20, 30, 50, 70, or 100 bases is used, and in yet other embodiments, an IPD averaged over all or substantially all of the template is used.

[0076] The sequence reads generated before and after the pause are analyzed within about a 20-, 30-, 50-, 70-, or 100-base window centered on the pause, and regions that flank the pause site and are complementary to one another are identified. Based upon the complementary sequences, their spacing, and other known factors that impact secondary structure formation (e.g., GC content, pH, salt concentration, and the like), the probability of a hairpin loop at that location in the template is determined. If this probability is high, the sequence reads flanking the pause site are re-examined to identify basecalls that do not match the complementarity of the hairpin, e.g., a non-complementary basecall or missing basecall within a stretch of complementary basecalls. Such non-complementary or missing basecalls have a higher probability of being errors than basecalls in the region that do not interrupt the complementarity between the regions upstream and downstream of the pause site. As such, the basecalls at these positions are reevaluated to determine if the initial basecall was erroneous. Further, knowledge of a given template's propensity for forming secondary structures that interfere with processivity of a polymerase can be used in future rounds of template-directed sequencing of the template to better call base positions in the vicinity of the interfering secondary structure, thereby improving accuracy of basecalls in the future rounds.

[0077] Additionally, since the duration of the pause is likely related to the strength of the secondary structure formed within the template, the duration can be used as a metric in determining the type, size, compositions, and likelihood of a secondary structure in the template molecule. In addition, for applications in which a single template is repeatedly subjected to template-directed synthesis, the replicate sequence reads that are generated are compared to one

another to determine if a given portion of the template consistently produces a pause in the synthesis reaction, which provides further evidence that the pause is due to the sequence context, e.g., secondary structure spontaneously forming in the template.

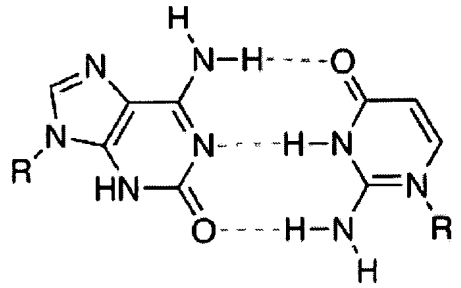
IV. Modified and Non-natural Nucleotide Analogs and Base Pairing

[0078] In certain aspects, methods, compositions, and systems are provided that utilize modified and/or non-natural nucleotide analogs and/or base pairing. For example, certain non-natural nucleotide analogs can be incorporated by a polymerase into a nascent strand opposite a modification, e.g., missing or damaged base. In certain embodiments, such non-natural nucleotide analogs are detectably labeled such that their incorporation can be distinguished from incorporation of a natural nucleotide or nucleotide analog, e.g., during template-directed nascent strand synthesis. This strategy allows real-time sequencing that generates reads that not only provide base sequence information for native bases in the template, but also modified bases without requiring further modifications to the standard methods (Eid, et al, *supra*). This method facilitates modification profiling in the absence of repeated sequencing of each DNA template, and is particularly well suited to *de novo* applications. In certain embodiments, the modified or non-natural nucleotide analogs are not incorporatable into the nascent strand and the polymerase can bypass the modification using a native nucleotide or nucleotide analog, which may or may not be labeled. Since the modified or non-natural analog has a higher affinity for the modification than a native analog, it will bind to the polymerase complex multiple times before a native analog is incorporated, resulting in multiple signals for a single incorporation event, and thereby increasing the likelihood of accurate detection of the modification. Similar methods for use in sequencing unmodified template nucleic acids are described in greater detail in U.S.S.N. 61/186,661, filed June 12, 2009 and incorporated herein by reference in its entirety for all purposes.

[0079] Since 5-MeC retains Watson-Crick hydrogen bonding with guanine, an incoming guanine nucleotide analog can be used to detect 5-MeC in the template strand. For example, a guanine nucleotide analog can be constructed to cross the major groove and sense the modified cytosine therein. In particular embodiments, a fused aromatic ring is linked to the N7 atom of the guanine of the nucleotide analog. This modified guanine nucleotide analog can "sense" the methyl group of 5-MeC and affect the base-pairing as compared to an unmodified guanine nucleotide analog. Such guanine nucleotide analogs are further described elsewhere, e.g., in International Application Pub. No. WO/2006/005064 and U.S. Patent No. 7,399,614. Similar modifications can be made to nucleotide analogs appropriate for SMRT™ sequencing applications, e.g., those with terminal-phosphate labels, e.g., as described in U.S. Patent Nos.

7,056,661 and 7,405,281; U.S. Patent Pub. Nos. 20070196846 and 20090246791; and U.S.S.N. 12/403,090, all of which are incorporated herein by reference in their entireties for all purposes. In certain embodiments, 5-MeC detection may be carried out using a modified guanine nucleotide analog described above that carries a detectable label that is distinguishable from detectable labels on other reaction components, e.g., other nucleotide analogs being incorporated. Such a strategy allows 5-MeC detection by observation of a signal, rather than or in addition to altered polymerase kinetics, which facilitates methylation profiling even in the absence of redundant or replicate sequencing of the template.

[0080] Certain embodiments use other non-natural base pairs that are orthogonal to the natural nucleobases pairs. For example, isoguanine (isoG) can be incorporated by a polymerase into DNA at sites complementary to isocytosine (isoC) or 5-methylisocytosine (^{Me}isoC), and vice versa, as shown by the following chemical structure and described in A.T. Krueger, et al., "Redesigning the Architecture of the Base Pair: Toward Biochemical and Biological Function of New Genetic Sets." *Chemistry & Biology* **2009**, *16*(3), 242, incorporated herein by reference in its entirety for all purposes.



Other non-natural base pairs that are orthogonal to the natural nucleobases pairs can also be used, e.g., Im-N^O/Im-O^N or A*/T* (described further in J. D. Ahle, et al., *Nucleic Acids Res* **2005**, *33*(10), 3176; A.T. Krueger, et al., *supra*; and A.T. Krueger, et al., Curr Opinions in Chem Biology 2007, *11*(6), 588).

[0081] In certain embodiments, a nucleic acid modification to be detected by the methods herein is 7,8-dihydro-8-oxoguanine ("8-oxoG") (also known as 8-oxo-7,8-dihydroguanine, 8-oxoguanine, and 8-hydroxyguanine). 8-oxoG is the major oxidative DNA lesion found in human tissue. Due to the relatively subtle modification to guanine in 8-oxoG, it may be bypassed by replicative DNA polymerases, which preferentially incorporate an adenine nucleotide into the nascent nucleic acid strand at the position where the complementary cytosine should be incorporated, thereby resulting in a mutation in the nascent strand (see, e.g., Hsu, et al. (2004)

Nature 431(7005): 217-21; and Hanes, et al. (2006) J. Biol. Chem. 281:36241-8, which are incorporated herein by reference in their entireties for all purposes). As well as introducing mutations *in vivo*, the bypass of such lesions by a polymerase during template-dependent sequencing reactions introduces errors into the sequence reads generated, and the presence of the damaged guanine nucleotide can also cause base misalignment, potentially adding further errors into a resulting sequence read. DNA synthesis opposite an 8-oxoG lesion has relatively very low specificity (kcat/Km) that is about 10^6-fold lower than incorporating a C opposite an unmodified G. See, e.g., Hsu, et al., *supra*. Further, due to its very low redox potential 8-oxoG can be more easily oxidized than unmodified guanine, and the 8-oxoG oxidation products are very effective blockers of DNA polymerases. See, e.g., Duarte, et al. (1999) Nucleic Acids Res 27(2):496-502; and Kornyushyna, et al. (2002) Biochemistry 41(51): 15304-14, the disclosures of which are incorporated herein by reference in their entireties for all purposes.

[0082] It has been shown that 8-oxoG alters both k_{cat} and K_m of steady-state incorporation kinetics, which are likely to cause altered pulse widths and IPD before incorporation of a nucleotide (G or A) into the complementary position in the nascent strand during template-directed sequencing reactions (see, e.g., Hsu, et al. and Hanes, et al., *supra*). These altered kinetic characteristics can be used to detect 8-oxoG in a template nucleic acid during real-time sequencing reactions. Further, a circular template that comprises both complementary strands of a region of interest (e.g., as described in U.S.S.N. 12/383,855 and 12/413,258, both filed March 27, 2009 and incorporated herein by reference in their entireties for all purposes) can be used to repeatedly sequence both strands of a region of interest, thereby generating redundant sequence information that can be analyzed to statistically determine how often a given position in the template has an A-G mismatch as compared to how often the correct base is incorporated at that position. The redundant sequence information increases the accuracy of correctly calling a position as a G or an 8-oxoG. For example, if the mismatch rate is 100%, then if one detects an A at the position, but then a G at the complementary position, then it is highly likely that the A detected was Hoogsteen base pairing with an 8-oxoG in the template. This strategy is similar to detection of 5-MeC modifications that have been deaminated to uracil prior to sequencing, as described in greater detail below.

[0083] The mismatch incorporation rate opposite 8-oxoG sites, as well as the degree to which IPD and pulse width are affected by 8-oxoG depend on the type of polymerase used in the reaction (see, e.g., Hsu, et al. and Hanes, et al., *supra*). As such, polymerase mutants can be designed to have increased kinetic sensitivity to 8-oxoG, or increased/decreased misincorporation rate opposite an 8-oxoG. Methods for designing polymerases for various embodiments of the invention are known in the art and provided elsewhere herein. Further,

multiple binding events are very likely at the site of modification, resulting in one or more signals not associated with incorporation into the nascent strand, and these multiple binding events can also occur at positions proximal to the modification, e.g., continuing for a few bases after the site of damage. These additional signaling events would provide a robust indicator of the site of modification. In addition, multiple sequencing reads for the region of the template comprising the modification are expected to contain variable numbers of extra signaling events at or proximal to the modification. As such, comparison of this redundant sequence data will also facilitate identification of loci comprising the modification.

V. Chemical Modification of Template

[0084] Direct detection of modifications (e.g., methylated bases as described above) without pre-treatment of the DNA sample, has many benefits. Alternatively or additionally, complementary techniques may be employed, such as the use of non-natural or modified nucleotide analogs and/or base pairing described elsewhere herein. In general, such complementary techniques serve to enhance the detection of the modification, e.g., by amplifying a signal indicative of the modification. Further, while the methods described herein focus primarily on detection of 5-MeC nucleotides, it will be clear to those of ordinary skill in the art that these methods can also be extended to detection of other types of nucleotide modifications or damage. In addition, since certain sequencing technologies (e.g., SMRT™ sequencing) do not require amplification of the template, e.g., by PCR, other chemical modifications of the 5-MeC can be employed to facilitate detection of these modified nucleotides in the template. For example, the difference in redox potential between normal cytosine and 5-MeC can be used to selectively oxidize 5-MeC and further distinguish it from the nonmethylated base. Such methods are further described elsewhere, and include halogen modification (S. Bareyt, et al., *Angew Chem Int Ed Engl* **2008**, *47*(1), 181) and selective osmium oxidation (A. Okamoto, *Nucleosides Nucleotides Nucleic Acids* **2007**, *26*(10-12), 1601; and K. Tanaka, et al., *J Am Chem Soc* **2007**, *129*(17), 5612), and these references are incorporated herein by reference in their entireties for all purposes.

Glycosylase Modification

[0085] By way of example, DNA glycosylases are a family of repair enzymes that excise altered, damaged, or mismatched nucleotide residues in DNA while leaving the sugar-phosphate backbone intact. Additional information on glycosylase mechanisms and structures is provided in the art, e.g., in A. K. McCullough, et al., *Annual Rev of Biochem* **1999**, *68*, 255. In particular, four DNA glycosylases (ROS1, DME, DML2, and DML3) have been indentified in *Arabidopsis thaliana* that remove methylated cytosine from double-stranded DNA, leaving an abasic site.

(See, e.g., S. K. Ooi, et al., *Cell* **2008**, *133*, 1145, incorporated herein by reference in its entirety for all purposes.) Furthermore, it has been shown that a 5'-triphosphate derivative of the pyrene nucleoside (dPTP) is efficiently and specifically inserted by certain DNA polymerases into abasic DNA sites through steric complementarity. (See, e.g., T. J. Matray, et al., *Nature* **1999**, *399*(6737), 704, incorporated herein by reference in its entirety for all purposes.)

[0086] In certain embodiments of single-molecule, five-color DNA methylation sequencing, DNA glycosylase activity can be combined with polymerase incorporation of a non-natural nucleotide analog (e.g., a pyrene analog (dPTP) as shown in Figure 4). For example, in certain embodiments, methylated cytosines are excised from a DNA sample treated with an *Arabidopsis* DNA glycosylase. Covalent linkage of a fifth fluorophore to the terminal phosphate of dPTP allows detection of abasic sites during polymerase-mediated DNA synthesis.

[0087] Figure 4 shows the principle of five-base DNA methylation sequencing. As shown in Figure 4A, genomic DNA is fragmented into pieces up to several kilobases in length, which serve as the DNA template. Figure 4B illustrates DNA glycosylase excising a 5-MeC from the template (black), leaving an abasic site. During SMRT™ sequencing, the DNA polymerase synthesizes the complementary strand and preferentially incorporates a fluorophore-phospholinked pyrene analog opposite the abasic site. This fluorophore has spectral characteristics distinct from those of the other four labeled nucleotides and indicates the presence of a 5-MeC in the original template. Further, error metrics can also be used to identify the modification, e.g., an increase in binding events for the pyrene analog may occur at the abasic site, as well as at downstream positions as the incorporated pyrene analog is "buried" in the nascent strand during subsequent incorporation events. In certain embodiments, a non-hydrolyzable pyrene analog carrying a detectable label is used at a concentration sufficient to bind (and be detected) several times at the abasic site before a hydrolyzable (and, preferably, distinctly labeled) analog is incorporated. Methods using non-hydrolyzable analogs are further described below.

[0088] A potential challenge in carrying out the above-described methods is that many DNA glycosylases display some lyase activity, e.g., bifunctional DNA glycosylase/AP lyases. These enzymes can cleave the phosphodiester backbone 3' to the AP (abasic) site generated by the glycosylase activity resulting in an abasic and unsaturated ribose derivative at that site, which could prevent a polymerase from incorporating the pyrene analog complementary to this site. In certain cases, it may be desirable to suppress any lyase activity of the *Arabidopsis* repair enzymes and enhance the desired glycosylase activity. Strategies for achieving this include site directed mutagenesis and the addition of a catalytically inactive AP endonuclease to the glycosylase reaction. (See, e.g., A. E. Vidal, et al., *Nucleic Acids Res* **2001**, *29*, 1285,

incorporated herein by reference in its entirety for all purposes.) A parallel protein mutagenesis program aims to enhance polymerase processivity in the presence of a dPTP analog. Other variations exploit ways in which the kinetics of pyrene incorporation into the abasic site are effected by fluorophore identity, the number of phosphates attached to the pyrene analog, and the structure of the linker connecting the fluorophore to the terminal phosphate group.

[0089] In other embodiments of single-molecule, five-color DNA methylation sequencing, DNA glycosylase activity can be combined with addition of a non-natural base to replace the methylated base. Briefly, after glycosylase-catalyzed excision of 5-MeC (with or without cleavage of the phosphodiester backbone), a class I or class II AP endonuclease is added to remove the abasic ribose derivative by cleavage at the phosphate groups 3' and 5' to the abasic site, thereby leaving 3'-OH and 5'-phosphate termini. A polymerase capable of extending from the free 3'-OH (e.g., Pol I or human pol β) and a non-natural base (e.g., isoC, isoG, or ^{Me}isoC) are added to incorporate the non-natural base into the abasic site. A DNA ligase (e.g., LigIII) is added to close the phosphodiester backbone by forming covalent phosphodiester bonds between the free 3'-OH and 5'-phosphates via ATP hydrolysis. Finally, a processive polymerase (e.g., Φ29 DNA polymerase) is used to synthesize a nascent nucleic acid strand complementary to the template strand, where the fifth nucleotide analog is the complement of the non-natural base that replaced 5-MeC in the template. For example, if the replacement base was isoC or ^{Me}isoC, then the fifth analog would be isoG. As such, the fifth analog would only incorporate into the nascent strand at positions complementary to 5-MeC sites in the template nucleic acid. In preferred embodiments, the fifth analog has a detectable label (e.g., fluorescent dye) that is distinct from labels on other reaction components, e.g, detectable labels on other nucleotide analogs in the reaction mixture.

[0090] Further, glycosylases exist or can be engineered for various DNA modifications, damage, or mismatches, so the methods described above are applicable not only for detection of 5-MeC, but also provide methods for detecting those other types of modifications, as well. Methods for the use of glycosylases for detection of other types of DNA damage are described in U.S.S.N. 61/186,661, filed June 12, 2009 and incorporated herein by reference in its entirety for all purposes. In certain embodiments, the pyrene (or similar) nucleotide analog can be non-hydrolyzable to increase the residence time and, therefore, lengthen the emitted signal indicative of the presence of the particular lesion of interest. A non-hydrolyzable fifth-base is eventually displaced by a hydrolysable analog and synthesis of the nascent strand continues. Alternatively, a fifth-base may be hydrolysable but may produce multiple separate signals prior to incorporation to increase the likelihood of detection.

Bisulfite Modification

[0091] In certain embodiments, the template may be modified by treatment with bisulfite. Bisulfite sequencing is a common method for analyzing CpG methylation patterns in DNA. Bisulfite treatment deaminates unmethylated cytosine in a single-stranded nucleic acid to form uracil (P. W. Laird, *Nat Rev Cancer* **2003**, *3*(4), 253; and H. Hayatsu, *Mutation Research* **2008**, *659*, 77, incorporated herein by reference in their entireties for all purposes). In contrast, the modified 5-MeC base is resistant to treatment with bisulfite. As such, pretreatment of template DNA with bisulfite will convert cytosines to uracils, and subsequent sequencing reads will contain guanine incorporations opposite 5-MeC nucleotides in the template and adenine incorporations opposite the uracil (previously unmethylated cytosine) nucleotides. If a nucleic acid to be treated with bisulfite is double-stranded, it is denatured prior to treatment. In conventional methods, amplification, e.g., PCR, typically precedes sequencing, which amplifies the modified nucleic acid, but does not preserve information about the complementary strand. In contrast, certain embodiments of the present invention include use of a template molecule comprising both strands of a double-stranded nucleic acid that can be converted to a single-stranded molecule, e.g., by adjusting pH, temperature, etc. Treatment of the single-stranded molecule with bisulfite is followed by single-molecule sequencing, and because the template retains both strands of the original nucleic acid, sequence information from both is generated. Comparison of the resulting sequence reads for each strand of the double-stranded nucleic acid will identify positions at which an unmethylated cytosine was converted to uracil in the original templates since the reads from the two templates will be non-complementary at that position (A-C mismatch). Likewise, reads from the two templates will be complementary at a cytosine position (G-C match) where the cytosine position was methylated in the original template. In certain preferred embodiments, a circular template is used, preferably having regions of internal complementarity that can hybridize to form a double-stranded region, e.g., as described in U.S. 12/383,855 and U.S. 12/413,258, both filed on March 27, 2009, and both incorporated herein by reference in their entireties for all purposes.

[0092] As described elsewhere herein, methylcytosine has an effect on IPD over a number of neighboring positions when compared to non-methylated cytosine. Uracil compared to thymine is like unmethylated cytosine compared to methylcytosine (i.e. the only difference between U and T is that T has an additional methyl group). Thus, the invention provides methods for performing bisulfite sequencing in which the polymerase kinetics (IPD and pulse width) or the mismatch incorporation rate are monitored in addition to the actual nucleotides being incorporated. Detection of a change in either of these kinetic parameters or in the mismatch rate

at the position in question, or at neighboring positions, is used to determine whether or not a position was always a T or is a U that was originally an unmethylated cytosine.

[0093] In certain embodiments, polymerase mutants are designed that are more sensitive to the difference between thymine and uracil in order to enhance the effect described above. Methods for designing polymerase variants are described in detail above and need not be repeated here.

[0094] Additionally or alternatively, PCR of uracil-containing oligonucleotides is not necessarily as efficient as PCR without uracil. This issue can bias the PCR amplification of bisulfite-converted DNA. The methods of sequencing-by-synthesis using bisulfite-modified templates described herein circumvent this problem by not using PCR amplification. However, the kinetics of these sequencing-by-synthesis reactions can be monitored to detect changes in kinetics due to the presence of uracil residues.

[0095] Further, the methods presented herein are useful for detecting PCR bias in the amplification of bisulfite-treated nucleic acids. For example, a few rounds of PCR could be performed on various oligos, some with uracil and some without (including controls with the same sequence but containing thymine in place of uracil). After performing sequencing-by-synthesis on all the resulting oligos, one could determine the percentage of oligos that still contain uracil. If it's different than the expected percentage given ideal (unbiased) PCR amplification, then a bias has been detected.

[0096] In yet further embodiments, a template nucleic acid is exposed to a reagent that transforms a modified nucleotide to a different nucleotide structure. For example, a bacterial cytosine methyl transferase converts 5-MeC to thymine (M. J. Yebra, et al., *Biochemistry* **1995**, *34*(45), 14752, incorporated herein by reference in its entirety for all purposes). Alternatively, the reagent may convert a methyl-cytosine to 5-hydroxy-methylcytosine, e.g., TET1 (M. Tahiliani, et al., *Science* **2009**, *324*(5929), 930, incorporated herein by reference in its entirety for all purposes). In further embodiments, the reagent may include a cytidine deaminase that converts methyl-cytosine to thymine (H. D. Morgan, et al., *J Biological Chem* **2004**, *279*, 52353, incorporated herein by reference in its entirety for all purposes). In yet further embodiments, a restriction enzyme that specifically alters a modification of interest can be used to create a lesion at the modification site. For example, DPNI cleaves at a recognition site comprising methyladenosine. Optionally, the cleaved template could be repaired during an analytical reaction by inclusion of a ligase enzyme in the reaction mixture. As noted elsewhere herein, nucleotides other than 5-MeC can also be modified and detected by the methods provided herein. For example, adenine can be converted to inosine through deamination, and this conversion affected by methylation of adenine, allowing differential treatment and detection of adenine and MeA.

DMS Modification

[0097] In certain embodiments, the template may be modified by treatment with dimethyl sulfate (DMS) prior to sequencing. DMS is a chemical that methylates the N7 position of guanine in dsDNA, and to a lesser extent the N3 position of adenine in dsDNA. If proteins are bound to a DNA treated with DMS, the proteins will block the methylation of the sequences to which they are bound. The bound proteins can then be removed and the DNA treated with piperidine, which breaks the DNA backbone by removal of the methylated bases. Protected regions of the DNA are identified as having been bound to the proteins during the DMS treatment. DMS also modifies the N3 position of cytosine and the N1 position of adenine in single-stranded DNA or RNA so these bases can no longer base pair with their complement. Since both these positions are involved in base-pairing, regions that are double-stranded during DMS treatment are protected from modification. Reverse transcriptase PCR and gel analysis is subsequently used to identify regions that were unmodified, and are therefore likely regions that adopt secondary structures that protect them from DMS treatment.

[0098] The present invention provides methods for real-time, single-molecule sequencing of nucleic acids that have been subjected to DMS treatment as a means for detecting both binding sites of nucleic acid binding agents, as well as sites of secondary structure formation, e.g., G-quadruplex structures (also known as G-tetrads or G_4-DNA; see, e.g., Zheng, et al. (2009) "Molecular crowding creates an essential environment for the formation of stable G-quadruplexes in long double-stranded DNA," Nuc Ac Res 1-12, incorporated herein by reference in its entirety for all purposes). For example, dsDNA bound to one or more nucleic acid binding agents is subjected to DMS treatment, and the binding agents are subsequently removed. The resulting dsDNA is subjected to template-directed sequencing and pulse metrics are monitored to identify locations where guanine or adenine were methylated. For example, A and G template nucleotides that cause a distinguishable change in one or more pulse metrics are identified as not having been bound by the agent(s), and A and G template nucleotides that do not cause a distinguishable change in one or more pulse metrics are identified as having been bound by the agent(s). In certain embodiments, the DMS treatment takes place *in vivo*, and the dsDNA is subsequently extracted and sequenced to study transcription factor binding in the cell. Alternatively, dsDNA can be extracted from cells and subsequently exposed to one or more nucleic acid binding agents prior to treatment with DMS *in vitro*. The DMS treatment can performed in solution, or can be performed after the dsDNA is immobilized, e.g., at a reaction site. The nucleic acid binding agents that can be studied include, but are not limited to, transcription factors, polymerases, ribosomes, and associated cofactors to DNA, and thereby study which DNA regions are being actively transcribed in different cells, in healthy vs.

diseased tissue, in different cell cycle stages, in response to various environmental stimuli, and the like. For example, in certain embodiments DMS is applied *in vivo* or *in vitro* to mRNAs bound by actively translating or stalled ribosomes. The resulting mRNA templates are subsequently sequenced in real time, and the reactions are monitored for altered kinetics, which are indicative of modified bases. Alternatively, the DMS-treated mRNAs can be heated to degrade modified regions, leaving only unmodified regions for sequencing. The sequence data so generated is used to identify the mRNAs to which a ribosome was bound, and therefore the mRNAs that were being actively translated in the sample from which they were extracted. Other methods of ribosome profiling are known in the art, e.g., Ingolia, et al. (2009) Science 324(5924):218-23, the disclosure of which is incorporated herein by reference in its entirety for all purposes.

[0099] In further embodiments, DNA and RNA secondary structure profiling can be performed by applying DMS to single-stranded DNA or RNA (e.g., mRNA, siRNA, microRNA, rRNA, tRNA, snRNA, etc.) and sequencing the DMS-modified nucleic acid using an appropriate polymerase. (Methods for sequencing RNA molecules using RNA dependent polymerases are described in detail in U.S.S.N 61/186,661, filed June 12, 2009 and incorporated herein by reference in its entirety for all purposes.) Regions of the treated nucleic acid that elicit altered polymerase kinetics are identified as regions that were single-stranded during the DMS treatment, and regions of the treated nucleic acid that do not elicit altered polymerase kinetics are identified as regions that were double-stranded during the DMS treatment and therefore likely contained duplex secondary structure, e.g., hairpins. In certain embodiments, the nucleic acid is heated prior to sequencing to cause degradation of the modified regions. The remaining, undegraded nucleic acid is subsequently subjected to sequencing and the sequence data so generated is used to identify regions of the original nucleic acid that formed secondary structures that prevent DMS modification.

[00100] DMS modification can also be used to map regions that form non-B-form secondary structures, some of which have regulatory roles *in vivo*. For example, G-quadruplexes consist of stacks of Gs that protect the guanosines from DMS-modification, even in the absence of a nucleic acid binding agent. Subsequent sequence analysis is used to identify regions that were protected from DMS modification, and therefore are likely to have had some protective secondary structure.

[00101] Further, although described primarily in terms of DMS modification, other types of chemical and/or enzymatic modifications can also be used in an analogous fashion, as will be clear to one of ordinary skill in the art based on the teachings herein. For example, other methods of DNA or RNA footprinting are particularly useful in the methods herein, including,

e.g., use of DNaseI, hydroxyl radicals, or UV irradiation for cleavage of nucleic acid that is not bound by an agent. Such methods are described more fully in the published literature.

[00102] The template altered by exposure to the reagent is sequenced, e.g., using a real-time, single-molecule methodology such as SMRT™ sequencing. In certain preferred embodiments, the sequencing is performed multiple times on the same template, e.g., by rolling-circle synthesis or another form of molecular redundant sequencing. The loci in the template containing altered nucleotides are identified by analysis of the resulting sequence reads. In cases in which the 5-MeC nucleotides were converted to non-altered nucleotide (e.g., thymine), molecular redundant sequencing on both the forward and reverse strands is useful for further refining the identification of the altered nucleotides since the transformation disrupts the normal Watson-Crick base pairing. For example, if MeC•G pair is converted to T•G, the forward and reverse reads will show non-complementary nucleotides at that position (A and C), indicating that the base pair in the template was non-standard, likely due to an alteration of a 5-MeC at that position. Methods for molecular redundant sequencing are further described in U.S. Patent No. 7,476,503 and U.S. Application Serial Nos. 12/383,855 (filed March 27, 2009), 12/413,258 (filed March 27, 2009), 12/413,226 (filed March 27, 2009), and 12/561,221 (filed September 16, 2009), all of which are incorporated herein by reference in their entireties for all purposes.

VI. Detection of Agent-Nucleic Acid Interactions

[00103] Another example of a biological process that may be monitored in accordance with the invention is association of a nucleic acid binding agent (e.g., a protein, nucleic acid, or small molecule) with a single nucleic acid molecule. As for the chemical modifications to the template described above, use of such agents serve to enhance the detection of the modification, e.g., by amplifying a signal indicative of the modification. Many types of agents bind to nucleic acids, such as transcription factors, RNA and DNA polymerases, reverse transcriptases, histones, nucleases, restriction enzymes, replication protein A (RPA), single-stranded binding protein (SSB), anti-DNA antibodies, DNA damage-binding agents, agents that bind altered nucleotides (e.g., methylated), small RNAs, microRNAs, drug targets, etc. In particular, transcription factors are involved in gene expression regulation and are thus very important for the study of diseases such as cancer. Further, RPA binds single-stranded DNA during replication to keep DNA unwound and accessible to the polymerase. Current technologies for detecting the binding of a protein transcription factor to a DNA molecule involve bulk detection. Certain aspects of the invention provide a method for detecting the binding of a transcription factor or other nucleic acid binding agent to a single molecule of DNA. The advantages of the methods described herein include, but are not limited to, improved resolution of kinetics (e.g., of

association and dissociation), binding loci, and statistical analysis; and greater sensitivity and simplicity.

[00104] In certain aspects, the invention provides detection of binding of a nucleic acid binding agent onto a single nucleic acid molecule through a technology that involves observing the activities of single molecules of polymerases in real time and with high multiplex capabilities, thereby allowing the screening of multiple nucleic acid binding agents (or other components of the reaction) with high throughput. In particular, the invention employs analogous processes used for single-molecule, real-time DNA sequencing, and with some modifications, exploits such processes to characterize various aspects of binding of nucleic acids by proteins of interest. Such sequencing technology has been previously described, e.g., in Eid, et al. (incorporated herein above). In certain preferred embodiments, one or more components of the reaction are immobilized at a reaction site, e.g., in an optical confinement such as a ZMW. Alternatively or additionally, multiple reactions can be simultaneously monitored by immobilizing them at discrete locations on a substrate, e.g. in an array of optical confinements. Further, to prevent displacement of the agent prior to a detectable affect on the reaction (e.g., a pause), the binding may be enhanced through various alterations to the reaction mixture (e.g., salt concentration, pH, temperature, etc.), or through alterations to the agent itself. For example, a DNA-binding protein may comprise various mutations that enhance binding under the conditions of the sequencing reaction, e.g., by lowering the Kd of the binding domain (e.g., a methyl binding domain) or by duplicating the domain to increase the effective concentration of the binding domain in the vicinity of the DNA template.

[00105] In certain preferred embodiments, a single nucleic acid template is bound to a sequencing engine (e.g., a polymerase or reverse transcriptase) that is synthesizing a nascent nucleic acid strand, e.g., during a template-directed sequencing reaction or a sequencing-by-synthesis reaction. The template can be any nucleic acid template appropriate for template-directed sequencing, e.g. single-stranded or double-stranded DNA, RNA, or a DNA/RNA hybrid. Further, the nucleic acid template can be linear or circular. For example, a dsDNA template can be bound by a polymerase in an optical confinement, e.g., a ZMW, as described above and in, e.g., Foquet, et al., and Levene, et al., both of which are incorporated herein supra. A nucleic acid binding agent, such as a transcription factor or DNA damage-binding agent, is added to the reaction mixture under conditions that promote binding of the agent to the template. If the template is bound by the agent in a location ahead of the polymerase, the bound agent impedes the translocation of the polymerase along the template, resulting in a pause or full stop in polymerization at or adjacent to the position at which the agent bound. Real-time monitoring of the ongoing sequencing reaction will allow detection of the pause or stop, which is indicative

of (i) the fact that the agent bound the template, and (ii) the position on the template that was bound by the agent, e.g., based on the sequence of nucleotides incorporated immediately prior to the pause or stop. Further, a consensus sequence for the binding site of the agent can be determined by statistical analysis of the "binding-affected" (e.g., containing a pause or truncated) sequence reads generated in the presence of the agent and the non-binding-affected (e.g., full-length) sequence reads generated in the absence of the agent. For example, truncated sequence reads (or sequence reads having detectable pauses) generated in the presence of the agent provide a location on the template at which the polymerase was blocked, and full-length reads generated in the absence of the agent provide the binding site sequence. In certain embodiments, sequence reads from the region of the template immediately downstream of the point at which the polymerase is blocked are analyzed together to find a sequence (specific or degenerate) they have in common, and this common sequence is identified as the consensus binding site for the agent. Such analyses are routine in nucleic acid sequence analysis and require no further elaboration here.

[00106] In certain embodiments, a nucleic acid binding protein of interest is introduced into a reaction mixture comprising a pool of nucleic acid templates. The pool of templates is exposed to the protein under conditions that promote binding, and polymerase enzymes are subsequently added to the reaction mixture and allowed to bind the templates, e.g., at a single-stranded region comprising a bound oligonucleotide primer. The reaction mixture further comprises a set of detectably labeled nucleotides, wherein each type of nucleotide in the set is linked to a distinct label that is optically identifiable during polymerization, thereby providing a distinct signal for each nucleotide incorporation event that identifies the base incorporated into the nascent strand. The polymerase-template complexes are immobilized on a substrate such that signals emitted from each complex are optically resolvable from signals emitted from every other complex on the substrate. Preferably, the reaction mixture is lacking a component required for polymerization to prevent polymerase activity prior to immobilization. Such a component is subsequently added to the reaction mixture allowing the polymerase to commence synthesis of a nucleic acid strand complementary to the template to which it is bound. For those templates that were not bound by the protein, synthesis continues unimpeded and the template is fully sequenced in the optical confinement, generating a full-length sequence read for the template. In contrast, the templates that were bound by the protein are processed by the polymerase until the bound protein is encountered on the template, at which time the polymerase will pause or stop polymerizing the complementary strand. The truncated sequence read generated from such a stalled polymerase-template complex will provide sequence information for the template upstream of the protein binding site. Statistical analysis of this sequence information, both at the

single molecule level and across the pool of templates, can be used to both identify the particular nucleic acid templates bound (or not bound) by the protein, as well as identifying the position at which the protein binds. For example, this technique can be used to map specific protein binding sites on the template, e.g., sequence-specific or lesion/damage-specific binding sites.

[00107] This assay can by easily modified to test the impact of various reaction conditions, e.g., pH, temperature, ion concentrations, and presence or absence of agents such as drugs, antibodies, or binding competitors. These tests can be used to identify optimal reaction conditions, e.g., for causing a pause or stop in an ongoing sequencing reaction or for binding to a particular subset of the pool of template nucleic acids. Further, the assay can be used to test variants and/or mutants of known nucleic acid binding proteins to screen such mutants for desired characteristics, such as binding under stringent conditions or having altered sequence specificity for binding. The assay can also be used to test variants and/or mutants of polymerase enzymes for desired characteristics, such as the ability to bypass a particular nucleic acid binding protein. Further, the specificity of binding can be explored by performing the assay with different pools of nucleic acid templates.

[00108] In certain embodiments, the nucleic acid binding protein is a transcription factor (TF) with a specific consensus binding sequence, e.g., TGACTCA for AP1 or GGACTTCC for NF-κB. DNA template molecules that contain the consensus binding sequence are bound by the TF at that sequence, and those that do not are not bound by the TF. When the translocating polymerase encounters a bound TF, the polymerase stops polymerizing, and the cessation of signals emitted from the complex is indicative that the TF bound the template and, therefore, that the template contains the consensus binding sequence. As noted above, various reaction conditions can be tested for their effect on either the binding of the TF or the ability of the polymerase to bypass it or displace it from the template. Statistical analysis of the sequence information from the DNA templates that were bound by the TF can be used to further characterize the TF, e.g., by (i) identifying genes targeted by the TF, e.g., using publicly available genome sequence data; (ii) identifying the consensus binding sequence, e.g., using sequence data generated from the same templates in the absence of the TF; (iii) studying the interaction of multiple transcription factors; (iv) modulation of TF binding by other proteins, small molecules, etc.; (v) testing the temperature sensitivity of binding; (vi) identifying and characterizing the abundance of particular DNA-binding proteins, e.g., in a cell extract; and the like. For example, the identity and abundance of DNA-binding proteins can be compared between a) different tissues, cell lines, cell developmental stages, species, or subspecies; b) healthy and diseased samples; and c) in the presence and absence of environmental stressors and/or various agents (e.g., drugs, toxins, etc.). Yet further, variants and mutants of different

components of the reaction mixture, e.g., TF, polymerase, template, etc., can be tested to identify those with particularly desirable characteristics, e.g., tight binding, protein displacement activity, non-consensus binding sequences with higher binding affinity to the TF, etc.

[00109] When implemented on in an arrayed format, such investigations would be highly parallel, enabling high-throughput screening assays. Arrays of reactions are carried out on highly multiplexed confocal fluorescence microscope systems (see, e.g., Lundquist, et al., incorporated herein above) in which the instrument detects fluorescent signals from each reaction site on the array, resulting in a highly parallel operation. Although preferred embodiments use arrays of zero mode waveguides, as described elsewhere herein, these assays could also be performed in other systems capable of real-time single-molecule detection, e.g., using total internal reflection fluorescence (TIRF) microscopy or waveguide technology.

[00110] Although certain embodiments are described in terms of nucleic acid binding proteins, it will be appreciated that the methods and systems described herein are equally applicable to other nucleic acid binding agents capable of pausing, stopping, or otherwise disrupting processive template-directed synthesis of a nascent nucleic acid molecule, e.g., nucleic acids and analogs and mimetics thereof (e.g., protein nucleic acids), lipids, sugar-oligoamides, intercalating dyes, major and minor groove binders, etc.

VII. Nucleic Acid Binding Agents as Analytical Tools

[00111] In certain aspects, nucleic acid binding agents are used in the methods, compositions, and systems of the invention to detect and/or reverse modifications in nucleic acid molecules. Such agents are typically used to enhance the response of a polymerase to a modification in the template nucleic acid. That is, the methods herein can be used to detect binding of an agent to the template, whether in response to a modification as described below, or simply an unmodified recognition site within the sequence of the template, as described above. Further, the effects of various agents on the creation, detection, or bypass of a nucleotide modification can also be tested and compared. For example, a template can be treated in various different ways (e.g., with and without a nucleic acid binding agent) and subsequently subjected to single-molecule sequencing-by-synthesis, which is monitored for a disruption in sequence read generation that is characteristic of binding of the agent to the template. In other embodiments, a template containing a known modification can be subjected to single-molecule sequencing-by-synthesis in the presence of various agents or reaction conditions. The reaction is monitored for the activity of the polymerase on the modified template to determine if the presence of any of the agents or other conditions impacts the ability of the polymerase to bypass or pause at the modification.

[00112] In certain specific embodiments, accentuating the differences in interpulse duration and/or pulse width between methylated and unmethylated DNA involves DNA binding proteins. It has been shown that some DNA polymerases stall when they encounter a DNA-bound protein complex. (See, e.g., M. Elias-Arnanz, et al., *EMBO J* **1997**, *16*, 5775, incorporated herein by reference in its entirety for all purposes.) In SMRT™ sequencing, this stall is detected as an unusually long interpulse duration that would end when the binding protein dissociates from the DNA template or is displaced by the translocating polymerase. There are a number of proteins that can bind stably and specifically to methylated DNA including members of the MBD family of human proteins, all of which contain a methyl-CpG binding domain (MBD). For example, MECP2, MBD1, MBD2, and MBD4 all bind specifically to methylated DNA, and are involved in repressing transcription from methylated gene promoters. Binding of these proteins to a template nucleic acid is expected to cause a translocating polymerase to pause proximal to the bound protein. As such, an increased pause duration during single-molecule sequencing reactions is indicative of a methylated base in the template nucleic acid. It is therefore important that the protein bind tightly to its target nucleic acid sequence. Natural MBD proteins only have micromolar Kd affinities for methyl-CpG sequences, so engineered MBD proteins that bind more tightly to the methylated template sequence can enhance detectability of methylated bases. For example, a multimerized MBD1 protein is provided in Jorgensen, et al., *Nucleic Acids Research* **2006**, 34(13), e96. Such engineered proteins can have a single methyl binding domain with a lower Kd (sub-micromolar) or multiple methyl-binding domains that increase the effective concentration of the methyl-binding domain in the vicinity of the methylated DNA template. More information on the MBD family of proteins is provided, e.g., in B. Hendrich, et al., *Mol Cell Biol* **1998**, 18(11), 6538; and I. Ohki, et al., *EMBO J* **2000**, 18(23), 6653.

[00113] In addition, the mammalian UHRF1 (ubiquitin-like, containing PHD and RING finger domains 1) protein binds tightly to methylated DNA and is required for its maintenance. Crystal structures of the SRA domain of this protein bound to DNA show that the 5-MeC is flipped out of the DNA duplex and stabilized by hydrophobic stacking and hydrogen bonding to SRA protein residues. (See, e.g., G. V. Avvakumov, et al. and H. Hashimoto, et al., both *supra*.) Finally, the monoclonal antibody to 5-MeC, used for methylated DNA immunoprecipitation, also binds specifically to methylated cytosine. (See, e.g., N. Rougier, et al., *Genes Dev* **1998**, *12*, 2108 ; and M. Weber, et al., *supra*, which are incorporated herein by reference in their entireties for all purposes.) All of the above-mentioned proteins are candidates for interfering with normal DNA polymerase processivity during SMRT™ sequencing. In order to enhance their polymerase stalling effects, these proteins can also be engineered to increase their affinity for

methylated DNA sites. See, e.g., H. F. Jorgensen, et al., *Nucleic Acids Res* **2006**, *34*, e96, the disclosure of which is incorporated by reference herein in its entirety for all purposes.

[00114] In yet further embodiments, an antibody against 5-MeC could be used to bind 5-MeC in a template nucleic acid, similar to the process used in methylated DNA immunoprecipitation assays (M. Weber, et al., *Nat Genet* **2005**, 37, 853). As such, the antibody essentially acts as an enhancer of the signal indicating the presence of the modification in the template by virtue of altering the polymerase dynamics. Various components of such reactions can be detectably labeled, e.g., the antibody, template, incorporated nucleotides, and combinations thereof, as described further elsewhere herein.

[00115] In still further embodiments, methyltransferases can be used to further facilitate detection of methyl-modified template nucleic acids. As described above, DNA methyltransferases catalyze the addition of methyl groups to DNA based upon recognition of methylation sites. For some methyltransferases (e.g., maintenance methyltransferases), the most active binding site in a nucleic acid is a hemi-methylated site in which one strand of the nucleic acid is methylated and the opposite strand is not. An enzymatically inactive methyl transferase (i.e., one that is unable to methylate nucleic acids) will therefore preferably bind to a hemi-methylated strand of DNA. In a real-time, template-directed sequencing reaction, a methylated single-stranded template becomes hemimethylated after nascent strand synthesis. A detectably labeled methyltransferase can therefore be detected interacting with the hemimethylated product of the synthesis reaction in real-time.

[00116] In certain embodiments, a circular template is used to permit rolling-circle synthesis by the polymerase in which a single-stranded circular template is converted to a double-stranded circular template. In preferred embodiments, the polymerase is capable of strand displacement such that after proceeding around the template once it begins to displace the nascent strand ahead of it as synthesis continues. This process eventually results in long concatemer containing multiple copies of the complement to the original template molecule. In such a system, a single-stranded methylated template is converted to a double-stranded hemimethylated template. A methyltransferase present in the reaction mixture can bind the hemimethylated sites and, if detectably labeled, this binding can be readily monitored in real time. When the polymerase encounters a bound methyltransferase, a pause may be detected prior to dissociation of the methyltransferase. The location of the pause in the resulting sequence reads can be used to map the position of the methylated site within the template molecule, even in the absence of a detectable label on the methyltransferase. For example, the pause can be used to identify the binding of the methyltransferase, e.g., in cases in which the methyltransferase is not detectably labeled, and in such cases the methyltransferase would essentially serve to extend

the pause at a methylated site, thereby facilitating identification of such a site in the template nucleic acid.

[00117] Other types of modifications can also be detected and/or reversed by nucleic acid binding agents. For example, among all types of DNA damage, oxidative base damage by reactive oxygen species (ROS) has been recognized as a major cause of cell death and mutagenesis in aerobic organisms (see, e.g., Finkel, et al. (2000) Nature 408(6809): 239-47, which is incorporated herein by reference in its entirety for all purposes). DNA oxidative lesions are primarily recognized and repaired by base excision repair (BER) pathways (see, e.g., Fromme et al. (2004) Adv Protein Chem 69: 1-41, which is incorporated herein by reference in its entirety for all purposes). In humans, the BER pathway for detecting and repairing a common oxidative lesion, 7,8-dihydro-8-oxoguanine ("8-oxoG"), begins with recognition of the lesion by a human oxoguanine DNA glycosylase 1 (hOgg1), which is a DNA glycosylase/apurinic (AP) lyase (see, e.g., Klungland, et al. (2007) DNA Repair (Amst) 6(4): 481-8, which is incorporated herein by reference in its entirety for all purposes).

[00118] The modified base 8-oxoG is discussed at length *supra*. Recent fluorescence and crystallography studies of hOgg1 found that this DNA glycosylase recognizes the oxidative DNA lesion 8-oxoG by scanning the DNA duplex, flipping the DNA base out, and transferring the damaged base from a pre-sampling binding site to the damage recognition binding site. Single-molecule experiments revealed the rapid sliding activity of hOgg1 on DNA duplex. For more detailed information on these studies, see Banerjee, et al. (2005) Nature 434(7033): 612-8, and Blainey, et al. (2006) Proc Natl Acad Sci USA 103(15): 5752-7, the disclosures of which are incorporated herein by reference in their entireties for all purposes.

[00119] In certain embodiments, the methods provided by the invention expose a nucleic acid template to a damage-recognition agent that binds to the template at a damaged nucleotide in a manner that blocks bypass of the lesion by a polymerase translocating along the template. The blockage causes a cessation of incorporation-dependent signaling from the reaction site, thereby indicating the damage-recognition agent has bound a damaged nucleotide in the template. In some aspects, the methods further include exposing a damaged template to additional reaction components that act to repair the damage, restoring the template and allowing dissociation of the damage-recognition agent from the previously damaged nucleotide. Elements of the damage-repair (e.g., base excision repair (BER)) machinery can be provided in the original reaction mixture, or can be added to an ongoing reaction. If the polymerase pauses but does not dissociate, the polymerization reaction can continue after DNA repair has been completed and the repair machinery has dissociated from the template or translocated away from the previously-damaged site.

[00120] In preferred embodiments, the damage-recognition agent is a protein involved in BER such as DNA gycosylases/apurinic (AP) lyases, e.g., hOGG1 (human oxoguanine DNA glycosylase 1), yOGG1 (yeast homolog of hOGG1), FPG protein (MutM; bacterial homolog of hOGG1); and others known in the art. Other proteins that can be used as a damage-recognition agent include other DNA glycosylases, e.g., AlkA, Nth, Nei, MutY, uracil DNA glycosylases (UDG), single-strand selective monofunctional uracil-DNA glycosylase (SMUG), thymine DNA glycosylase (TDG), NEIL (e.g., hNEIL1 and hNEIL2), etc. Reaction components for repair of a damaged template bound by the damage-recognition agent include, e.g., AP endonucleases, DNA polymerase beta, and ligase, among others known in the art. See, e.g., McCullough, et al. (1999) Annu Rev Biochem 68: 255-85, which is incorporated herein by reference in its entirety for all purposes. Further, additional proteins that stimulate damage recognition may also be included in an analytical reaction; e.g., HAP1 (APE1) protein has been found to stimulate hOGG1 activity (Vidal, et al. (2001) Nuc. Ac. Res. 29(6):1285-1292).

[00121] In certain embodiments, more than one polymerase may be present in a template-directed sequencing reaction in which one or more lesions may be present on the template nucleic acid. For example, "bypass polymerases" have been discovered in both prokaryotes and eukaryotes, most of which belong to the Y-family of polymerases and/or are considered to be repair polymerases. In contrast to replicative polymerases, they operate at low speed, low fidelity, and low processivity. However, because their active sites adopt a more open configuration than replicative polymerases they are less stringent and can accommodate altered bases in their active sites. For more information on bypass polymerases, see, e.g., Cordonnier, et al. (1999) Mol Cell Biol 19(3):2206-11; Friedberg, et al. (2005) Nat Rev Mol Cell Biol 6(12):943-53; Holmquist, et al. (2002) Mutat Res 510(1-2):1-7; Lehmann, A.R. (2002) Mutat Res 509(1-2):23-34; Lehmann, A.R. (2006) Exp Cell Res 312(14):2673-6; Masutani, et al. (1999) Nature 399(6737):700-4; and Ohmori, et al. (2001) Mol Cell 8(1):7-8, the disclosures of which are incorporated herein by reference in their entireties for all purposes. Certain of these polymerases can bypass lesions in a nucleic acid template and carry out "translesion synthesis" or TLS. As such, DNA replication in the presence of such lesions was found to require multiple polymerases and the "polymerase switch model" was developed (see, e.g., Friedberg, et al. (2005) Nat Rev Mol Cell Biol 6(12):943-53; Kannouche, et al. (2004) Cell Cycle 3(8):1011-3; Kannouche, et al. (2004) Mol Cell 14(4):491-500; and Lehmann, et al. (2007) DNA Repair (Amst) 6(7):891-9, all of which are incorporated herein by reference in their entireties for all purposes). In brief, the polymerase switch model is model for lesion bypass during replication that involves replacement of a replicative polymerase with a bypass polymerase at a lesion, synthesis of the nascent strand by the bypass polymerase until past the lesion, and subsequent

replacement of the bypass polymerase with the more processive, higher fidelity replicative polymerase for continued replication past the lesion.

[00122] In certain preferred embodiments, one or more bypass polymerases is included in a template-directed nucleic acid sequencing reaction. For example, during the course of a reaction in which a replicative polymerase encounters and is blocked by a lesion in a template nucleic acid, the replicative polymerase is replaced by a bypass polymerase at the site of the lesion, and the bypass polymerase synthesizes a segment of the nascent strand that is capable of base-pairing with the damaged base, and may further include one or more bases prior to and/or past the site of the lesion in a process called "translesion synthesis." The limited processivity of the bypass polymerase causes it to dissociate and be replaced by the replicative polymerase following translesion synthesis. The replicative polymerase continues to synthesize the nascent strand until another blocking lesion is encountered in the template, at which point it is once again replaced by a bypass polymerase for translesion synthesis. (See, e.g., Friedberg, et al. (2005) Nat Rev Mol Cell Biol 6(12):943-53; and Kannouche, et al. (2004) Mol Cell 14(4):491-500, incorporated herein by reference above.) The process continues until the template has been replicated or the reaction is terminated, e.g., by the investigatior. One particular advantage of the polymerase switch method of template-dependent sequencing is that is it tolerant of most types of lesions in the template nucleic acid. As such the damaged template can be sequenced through a lesion, thereby allowing reinitiation of synthesis downstream of the lesion and increasing read lengths on lesion-containing templates.

[00123] Various different bypass polymerases known to those of ordinary skill in the art can be used with the methods and compositions provided herein, include prokaryotic polymerases (e.g., DNA polymerase IV, polymerase V, Dpo4, Dbh, and UmuC) and eukaryotic polymerases (e.g., DNA polymerase η, DNA polymerase ι, DNA polymerase κ, and Rev1). In eukaryotes, multiple bypass polymerases participate in translesion synthesis, and a processivity factor, proliferating cell nuclear antigen ("PCNA"), is also required and can be included in a sequencing reaction.

[00124] In certain preferred embodiments, the template or primer is immobilized during the template-dependent synthesis reaction to ensure that the template remains at the reaction site during polymerase switching. Alternatively or additionally, one or more polymerases can be immobilized at the reaction site. Various immobilization strategies useful in different aspects of the invention are provided elsewhere herein.

[00125] Since the portion of the nascent strand corresponding to the site of the lesion in the template is synthesized by a bypass polymerase, the sequence reads generated therefrom are expected to be less reliable than those generated from regions of the nascent strand synthesized

by the replicative polymerase. As such, generation of redundant sequence information during a sequencing reaction is a preferred means of generating complete and accurate sequence reads. Redundancy can be achieved in various ways described elsewhere herein, including carrying out multiple sequencing reactions using the same original template with the sequence data generated in the multiple reactions combined and subjected to statistical analysis to determine a consensus sequence for the template. For example, the sequence data from a region in a first copy of the template that was replicated by a lower fidelity bypass polymerase can be supplemented and/or corrected with sequence data from the same region in a second copy of the template that was replicated with a higher fidelity replicative polymerase. Further, a template can be amplified (e.g., via rolling circle amplification) to generate a concatemer comprising multiple copies of the template that is subsequently sequenced to generate, a sequencing read that is internally redundant. The sequence data from a first segment of the concatemer (corresponding to a first region of the template) that was replicated by the bypass polymerase can be supplemented and/or corrected with sequence data from a second segment of the concatemer (that also corresponds to the first region of the template) that was replicated by the replicative polymerase. Further, as noted above, redundancy can also benefit identification and characterization of lesions that occur in the same position in a plurality of templates, or that occur at a single position in a template that is subjected to resequencing. For example, since base incorporation by the bypass polymerase is promiscuous, replicate sequencing reads for the region containing the lesion may show more than one "complementary base" being incorporated at the same position in different reads (of the same or an identical template), and detection of such promiscuity is indicative that there is a lesion at that position in the template nucleic acid(s).

[00126] In certain embodiments, a polymerase in the reaction mixture may comprise a detectable label to indicate when that polymerase is associated with the template nucleic acid. For example, a bypass polymerase can comprise a detectable label that will indicate when the bypass polymerase is carrying out translesion synthesis. The nucleotides incorporated into the nascent strand during that time can therefore be identified and "tagged" as corresponding to a region of the template that contains one or more lesions, thereby allowing targeting of statistical analysis to these sequence reads, e.g., as described above.

[00127] In yet further embodiments, a nucleic acid binding agent specifically binds to secondary structure in the nucleic acid template, e.g., hairpin loops, stem-loops, internal loops, bulges, pseudoknots, base-triples, and the like. Binding of an agent to such structures inhibits passage of the polymerase through the structures to a greater extent than the enzyme is inhibited in the absence of the agent, thereby increasing the resulting pause time and facilitating detection of the secondary structure. Examples of agents that have binding specificity for specific

structures and/or strandedness in nucleic acids include, e.g., intercalating agents, nuclease-deficient endonucleases (e.g., with a specificity for a double-stranded region within a stem-loop structure), polymerases, and various eukaryotic initiator proteins.

[00128] As noted above, various different types of templates for template-directed polymerization reactions can be used, e.g., single-stranded or double-stranded DNA, single-stranded or double-stranded RNA, and analogs and mimetics thereof. Further, the template can contain a combination of single-stranded and double-stranded regions, e.g., such as the templates described in U.S.S.N. 12/383,855 and 12/413,258, both filed on March 27, 2009 and incorporated herein by reference in their entireties for all purposes. The type of template used is limited only by the substrate specificity of the polymerase and damage-binding agent in the reaction. For example, Figure 5 provides an illustrative embodiment of such a reaction comprising a linear template and a damage-binding agent that recognizes a lesion in a single-stranded template. In A, the damage-binding agent (305) is scanning a linear, single-stranded nucleic acid template (310) ahead of a polymerase (315) performing template-directed polymerization of a nascent nucleic acid strand (320). In B, the damage-binding agent (305) has detected and bound to a lesion (325) in the single-stranded template (310). In C, the polymerase (315) has caught up with the damage-binding agent (305) and its progress along the template (310) is blocked. In D, the lesion has been repaired by repair machinery (330) recruited by the damage-binding agent (305). In E, the repair machinery has dissociated from the template (310) and the damage-binding agent (305) has translocated away from the previously damaged site, thereby allowing the polymerase (315) to resume synthesis of the nascent strand (320).

[00129] In some embodiments, a damage-binding agent with specificity for double-stranded nucleic acid may be used in a reaction comprising a single-stranded template, e.g., when the scanning and damage detection/binding is expected to occur after the polymerase has converted the single-stranded template to a double-stranded template by template-dependent polymerization, e.g., after a single-stranded circle has been converted to a double-stranded circle during "rolling-circle replication." For example, although an initial substrate in a reaction is a circular single-stranded nucleic acid template, after a polymerase has processed the template one time it becomes a double-stranded template and an appropriate substrate for a damage-binding agent that specifically scans and binds double-stranded nucleic acid. For example, Figure 6 illustrates an embodiment comprising a circular template and a damage-binding agent that recognizes a lesion in a double-stranded template. In A, the damage-binding agent (405) is scanning a circular, double-stranded nucleic acid template (435) ahead of a polymerase (415) performing template-directed polymerization while displacing the 5' end of the nascent nucleic acid strand being synthesized (440). In B, the damage-binding agent (405) has detected and

bound to a lesion (425) in the double-stranded template (435), and the progress of the polymerase (415) is blocked by the bound damage-binding agent (405). In C, the lesion has been repaired and the damage-binding agent (405) has translocated away from the previously damaged site, thereby allowing the polymerase (415) to resume synthesis of the nascent strand (420).

[00130] Although various embodiments are described in terms of recognition and, optionally repair of 8-oxoG lesions, other types of DNA damage can also be addressed by the methods herein. For example, in the case of hOGG1, the *N*-glycosylase activity releases damaged purines from double-stranded DNA, generating an apurinic (AP) site. The AP-lyase activity cleaves 3′ to the AP site leaving a 5′ phosphate and a 3′-phospho-α, β-unsaturated aldehyde. In addition to 8-oxoG (when paired with cytosine), hOGG1 also recognizes and removes 8-oxoA (when base paired with cytosine), foramidopyrimidine (fapy)-guanine and methy-fapy-guanine (Bjoras, M. et al. (1997) *EMBO J.*, 16, 6314-6322; and Boiteux, S. and Radicella, J. (1999) *Biochimie*, 81, 59-67, the disclosures of which are incorporated herein by reference in their entireties for all purposes). Other types of DNA damage that can be bound and, optionally, repaired by the methods herein include BER enzymes that repair other DNA base lesions (small DNA base modification), e.g. AAG/MPG for methylated lesions, UDG/SUMG1 for repairing uracil in DNA, APE for abasic sites, etc. Also included are nucleotide excision repair (NER) enzymes that repair more bulky DNA lesions, such as DNA base adducts and DNA intra- and inter-strand crosslinks. Furthermore, although the DNA polymerase switch methods described above are suitable for detecting and bypassing most DNA lesions that block a replicative polymerase, certain small base modifications like 8-oxoG can be bypassed by a replicative polymerase, and thus methods that include binding agents that block the polymerase at the site of a lesion can help ensure that such lesions are detected, and optionally removed, from the template to prevent the sequence data generated from the template-dependent sequencing reactions to be adversely affected.

[00131] In certain embodiments, hOGG1 is included in a template-directed DNA sequencing reaction in the presence of a polymerase and a set of nucleotides, each of which bears a label that is optically detectable and that distinctively identifies the base (e.g., A, G, T, or C). Detection of an optical signal upon interaction with the polymerase and incorporation into the nascent strand allows the practitioner to identify the base incorporated and, by complementarity, the sequence of the template DNA molecule. In preferred embodiments, the incorporation of nucleotides into the nascent strand continues in a processive fashion, generating an ordered set of optical signals that can be analyzed to provide a sequence for both the nascent strand and, by complementarity, the template strand. The hOGG1 enzyme associates with the

template, "scans" for damage, and specifically binds to locations at which such damage occurs. As such, if the template DNA molecule contains or acquires (e.g. during the course of the analytical reaction) DNA damage recognized by hOGG1, it is bound by hOGG1, bypass of the lesion by the polymerase is blocked, and the incorporation-based signal is slowed or stopped (e.g., by stalling or dissociation of the polymerase). Although such a blockage can cause dissociation of the polymerase, in certain preferred embodiments the polymerase merely pauses until the damaged nucleotide is repaired and hOGG1 and any other repair machinery dissociates from the template, at which time polymerization resumes and additional sequence data is generated from the template at and downstream of the site of the previously damaged nucleotide.

[00132] In certain preferred embodiments, one or more reaction components is immobilized at a reaction site, e.g., in an optical confinement such as a zero mode waveguide (ZMW). In some embodiments, the polymerase is immobilized and the nucleic acid template and damage-binding agent are free in solution. Methods for immobilizing a polymerase enzyme are available in the art and provided elsewhere herein. In other embodiments, the nucleic acid template can be immobilized at the reaction site with the polymerase and damage-binding agent free in solution. For example, in preferred embodiments the damage-binding agent translocates upon the template faster than the polymerase so it does not impede progress of the template-dependent sequencing reaction on an undamaged template. However, upon binding a lesion, the damage-binding agent will stop and bind to the site, blocking progress of a translocating polymerase past the lesion. For example, hOGG1 translocates much faster than phi29 polymerase on undamaged DNA, but after encountering a damaged nucleotide the enzyme will bind to the site and wait for other components of the BER machinery. Alternatively or additionally, the damage-binding agent may be immobilized at the reaction site. For example, in the case of hOGG1 only a single enzyme is required for DNA binding, scanning, and lesion recognition. Immobilization of a single damage-binding agent at the reaction site increases the likelihood that a single template at each reaction site will be scanned for damage. Methods for immobilizing various reaction components are known in the art as described elsewhere herein.

[00133] In certain aspects, the methods for detection of nucleic acid damage can be used to test various elements of an experimental system to identify sources of such damage. For example, various buffer conditions or other components of an analytical reaction (e.g., reaction components or radiation that can induce production of oxygen radicals) can be tested to identify those that cause the least amount of damage for use in an experimental system. Further, such damage can be intentionally introduced into a nucleic acid template by the practitioner, e.g., at one or more specific locations in a template. This provides a means for controlling the progress of the polymerase, and therefore controlling the timing of production of sequence reads from

different portions of the template. For example, if the template is extremely long (e.g., thousands or tens of thousands of base pairs in length), it may be beneficial to temporarily pause the reaction at one or more points on the template to allow orientation of the sequence read to the template. In particular, a pause in emission of signal pulses is indicative that the polymerase has reached a particular location on the template, and the investigator can reinitiate polymerization by addition of repair agents/proteins to the reaction mixture. Such repair agents may be washed out of the reaction mixture and, optionally, reintroduced at a later point during the course of the reaction, e.g., by buffer exchange.

VIII. Data analysis

[00134] Analysis of the data generated by the methods described herein is generally performed using software and/or statistical algorithms that perform various data conversions, e.g., conversion of signal emissions into basecalls, conversion of basecalls into consensus sequences for a nucleic acid template, and conversion of various aspects of the basecalls and/or consensus sequence to derive a reliability metric for the resulting values. Such software, statistical algorithms, and use thereof are described in detail, e.g., in U.S. Patent Publication No. 20090024331 and U.S.S.N. 61/116,439, the disclosures of which are incorporated herein by reference in their entireties for all purposes. Specific methods for discerning altered nucleotides in a template nucleic acid are provided in U.S.S.N. 61/201,551, filed December 11, 2008, and incorporated herein by reference in its entirety for all purposes. These methods include use of statistical classification algorithms that analyze the signal from a single-molecule sequencing technology and detect significant changes in one or more aspects of signal morphology, variation of reaction conditions, and adjustment of data collection parameters to increase sensitivity to changes in signal due to the presence of modified or damaged nucleotides.

[00135] In certain aspects, the invention provides methods for detecting changes in the kinetics (e.g., slowing or pausing) or other reaction data for real-time DNA sequencing. As discussed at length above, detection of a change in such sequencing applications can be indicative of secondary structure in the template, the presence of modifications in the template, the presence of an agent bound to the template, and the like. It is appreciated that the kinetic activity of single molecules does not follow the regular and simple picture implied by traditional chemical kinetics, a view dominated by single-rate exponentials and the smooth results of ensemble averaging. In a large multi-dimensional molecular system, such as the polymerase-DNA complex, there are processes taking place on many different time scales, and the resultant kinetic picture can be quite complex at the molecular level. (See, e.g., Herbert, et al. (2008) Ann Rev Biochem 77:149.) As such, a real-time single-molecule sequencing technology should be

adaptable to such non-exponential behavior. For example, pauses during a real-time sequencing reaction are detectable as regions in the trace of observed signals over time in which it appears that the enzyme has significantly slowed as compared to the average rate of incorporation. As such, methods are provided to analyze the data generated in the vicinity of a pause site, and in particular algorithmic methods for classifying and removing or down-weighting the occurrence of pauses in the context of single-molecule sequencing. General information on algorithms for use in sequence analysis can be found, e.g., in Braun, et al. (1998) Statist Sci 13:142; and Durbin, et al. (1998) *Biological sequence analysis: Probabilistic models of proteins and nucleic acids*, Cambridge University Press: Cambridge, UK.

[00136] In certain preferred embodiments, the methods utilize a segmentation algorithm for discriminating pause regions in a real-time signal generated by monitoring single-molecule kinetics, in particular by monitoring DNA synthesis by DNA polymerase. The central observation is that during a pause the density of signal events (incorporations) is lowered, where the density refers to the number of events per a fixed unit of time. At the same time, stochastic events arising from Poisson processes, such as sticks (signals that do not correspond to an incorporation event, e.g., dyes that enter the detection volume but are not linked to nucleotides that are incorporated into the nascent strand) should continue at the same density as normally observed. Figure 7 illustrates an observation of true incorporations (solid line) versus stochastic pulses (dashed line) across time. A pause is identified in the region of the trace in which the observation of true incorporations dips below the observation of stochastic pulses. As such, by observation of differences in local densities, pauses in incorporation activity can be identified.

[00137] Other features that are related to pausing and can contribute to a full model of the phenomenon. In particular, the local sequence context of the template strand can also influence and inhibit the activity of the polymerase along the template. For example, the local sequence context that influences and/or inhibits activity of the polymerase may extend for at least about one, two, three, four, five, seven, ten, fifteen, twenty nucleotide positions, and these positions may lie upstream or downstream of the modification, or may flank the modification in the template. Other known models can also be used in the methods described herein, as will be clear to one of ordinary skill upon review of the teachings herein.

[00138] In a preferred embodiment, an algorithm for use in detecting changes during template-directed nucleic acid synthesis comprises the following general steps. First, a classifier is created that can distinguish between true incorporations and stochastic pulses. Features that can help discriminate between the two include, e.g., pulse height, pulse width, local signal-to-noise ratio, dye channel, and the χ^2 metric for the measured spectrum. Many different statistical classification algorithms known in the art can be used in this classifier. Certain preferred

algorithms include classification-and-regression trees (CART), naïve Bayesian classifiers, kernel density methods, linear discriminant functions, and neural networks. Further, the pulse classifier does not need to be particularly powerful (in an optimal specificity/sensitivity sense), because a strength of the approach relies on the greater significance associated with observing clusters of weakly significant events.

[00139] A second step is to slide a fixed-length window across the observed signal trace and count the number of incorporations versus stochastic pulses in each window using the classifier. Choice of the window size is determined by the length scale of events to be detected; a reasonable choice in practice is 10 seconds, but a practioner may increase or decrease the window size according to a particular implementation of the invention. Regions of the trace in which the stochastic pulse density exceeds the incorporation density for an extended period of time, e.g., 5-15 seconds, or more preferably about 10 seconds, are identified as corresponding to a likely pause site. These regions can be found using standard peak-finding techniques, e.g., threshold detection, finite-state machines, multi-scale methods, etc. This method has the further advantage of focusing on pause regions that have detrimental effects in the downstream use of sequencing data. The mere occurrence of a pause during sequencing is not consequential to the use of the data for DNA sequence analysis, however the occurrence of a large number of stochastic pulses in the pause region does complicate the use of the resulting data.

[00140] A variation on this exemplary algorithm is to use the time between true pulses that are identified by the classifier as a discriminator for finding pause regions, where regions that have a large difference between true pulses are candidate regions for pauses. For example, A plot of Δt (the time between true pulses) versus time will have local maxima at the location of candidate pauses.

[00141] A more sophisticated algorithm for use in detecting pause regions is a segmenting algorithm based on a hidden Markov model (HMM) architecture. Figure 8 provides an illustrative example of a simple hidden Markov model for classifying pause (P) versus sequencing (S) states within a sequencing trace. The use of this model assumes that each pulse can be labeled as either a probable incorporation (A_S, C_S, G_S, T_S) or a stochastic pulse (A_P, C_P, G_P, T_P). By fitting this model on multiple instances of sequence data (using e.g. the Baum-Welch algorithm), good emission and transition probabilities that correspond to the hidden pause and sequencing states can be generated. When subsequently presented with any particular signal (observed labels), the model can be queried for the underlying sequencing of hidden states using the Viterbi algorithm. This model is more powerful than the more simple algorithmic approaches suggested above in several ways. First, this model permits the modeling of per-nucleotide likelihoods for incorporations or sticks during the pauses or sequencing states. An

example where this is useful is if a stick/incorporation classifier for one nucleotide is particularly effective and if the pulses for another nucleotide are difficult to classify in this way. This model permits some nucleotide-specific differences in the classification power for stochastic pulses versus incorporations. A further advantage of this approach is that it is more adaptable to detecting regions across multiple time scales, where HMM segmentation approaches are usually better able to handle multi-time scale classification. The final assignment of pause regions is made by computing the log-odds ratio $\log \frac{p\langle P|x_i\rangle}{p\langle S|x_i\rangle}$ across the pulses (x_i) and identifying regions of high pause likelihood.

[00142] A more powerful algorithmic architecture for segmentation is the use of the conditional random field framework (CRF). The object is to predict the conditional probability of a signal arising from a pause or sequencing state given the observed pulses: $p\langle y|x\rangle = \frac{\exp[w^T F(x,y)]}{\sum_{y'} \exp[w^T F(x,y')]}$, where y is the sequence of desired labels (pause, sequencing), x is the observed pulse data (both basecalls and other pulse features), w is the weight vector learned from the training data, and the F function is the feature vector. The weights in the CRF can be trained using labeled sequences using standard techniques from the CRF literature (for example, Lafferty et al (2001) Proc. 18th International Conference on Machine Learning, 282-289, which is incorporated herein by reference in its entirety for all purposes). By labeling the sequence in this way, regions of high pause probability can be identified using the methods described above. One advantage of this method is the lack of a requirement for a per-pulse classifier for distinguishing between incorporation and stochastic pulses. It can also better integrate knowledge of the inter-pulse spacing and other information, such as sequence context, into a broad model. Potential disadvantages are the large amount of training data required to build the model and the algorithmic complexity involved in constructing a CRF model.

[00143] The application of such algorithmic methods to identify pause sites and/or regions with locally high stochastic pulses in sequencing trace data is useful in a number of contexts. For example, pulses in regions that are predicted to exhibit enzyme pausing can be labeled as less confident (lower quality value) for their use in downstream analyses such as sequence variant detection in resequencing applications or overlap detection for *de novo* assembly. In other embodiments, pulses from regions with a high probability of containing a cluster of stochastic pulses can be removed from the reported basecalled sequence, thereby improving the accuracy of sequence data for downstream use without resorting to secondary information such as quality values. In other embodiments, the occurrence of pauses can be

associated with other observables of interest, such as the probable DNA sequence or the occurrence of modified nucleotide bases. For example, sequences upstream of a pause site can be called in part based on their known effect on pausing. That is, if a pause occurs downstream of a sequence, then the sequence is more likely to be one that facilitates or exacerbates pausing than one that has no effect or that reduces the likelihood of pausing. As such, if sequencing of a modification is known to increase the likelihood of pausing, then this information can be incorporated into a Bayesian likelihood model for identifying modified bases. In further embodiments, the pause detection methods described herein can also be used to increase the understanding of the biophysics of polymerase activity, thereby providing useful feedback to efforts to better develop single-molecule, real-time sequencing techniques.

[00144] Algorithms for the identification of regions in sequence data belong to the general category of sequence labeling or segmentation algorithms, which are generally known in the art. The mapping of this problem to sliding-window analysis, HMMs, or CRFs is natural in this context. Other algorithms that approach the same problem are multiple change-point analysis such as the Gibbs sampler (see, e.g., Lee, P.M. (2004) *Bayesian Statistics: An Introduction*, Oxford University Press: New York, NY, the disclosure of which is incorporated herein by reference in its entirety for all purposes), or locally weighted polynomial regression (see, e.g., Braun, et al., *supra*).

[00145] In general, data analysis methods benefit when the sequencing technology generates redundant sequence data for a given template molecule, e.g. by molecular redundant sequencing as described above. The distribution of IPDs for each read at that position is an exponential. The decay constant for the exponential of a methylated base and for that of an unmethylated base may be different. However, because of the large amount of overlap between two exponentials, it is still challenging to use one read to distinguish between the two populations. However, if one takes the mean of multiple reads at a single position, the distribution of this mean is a gamma function (convolution of several exponentials), which is more Gaussian-like and better separated than exponentials. This enables better distinguishability of the two populations. For example, Figure 13 provides actual data showing that for two different positions in a single circular template, one always unmethylated, and one differentially methylated, an increase in the number of reads for the template corresponds to an increased resolution between IPDs for methylated vs. unmethylated adenosines. If the underlying distributions are exponential, as just discussed, then the mean value is the only metric that can be used for making the distinction (the standard deviation is the same as the mean). If the distribution is non-exponential for each read position, as it would be for the methylcytosine IPD' that is weighted over numerous neighboring positions and thus itself has a gamma-like

distribution, then when doing consensus reads of the same position, one can take into account the mean of the gamma-like weighted IPD' distributions along with other information, e.g. its standard deviation, its skewness, or other characteristics of the distribution. Figure 10 shows actual molecular consensus distributions for methylcytosine, given the underlying gamma-like weighted IPD' distributions of individual reads, but in this figure only the means of these underlying distributions were utilized. The plotted distributions could become even more well-separated if other characteristics had been taken into account. The data used to generate Figures 13 and 10 is more fully described in the Examples herein.

[00146] In certain embodiments, methods may be employed that use weighted sums of signal features at multiple positions to determine the status of a base, e.g., whether or not it is methylated in a template nucleic acid. In particular, interpulse duration (IPD) information from multiple positions can be used to determine whether or not a given cytosine is methylated, e.g., by comparing nascent strand synthesis data for a differentially methylated template (Me+) to such data for a fully unmethylated template (Me-). In certain preferred embodiments, a pseudo IPD is created for a given template position that is actually a weighted sum of the IPDs for the surrounding positions. More specifically, $IPD'_j = \sum_i w_i \times \frac{IPD_{ji}}{\langle IPD_j \rangle_{Me-}}$, where j is the index of the cytosine in question; i is an index that ranges over all the neighboring positions that yield a change in IPD due to cytosine being either methylated or unmethylated; and $\langle IPD_i \rangle_{Me-}$ is the average IPD at that particular position in the Me- template. The individual weights for the multiple positions (all together which would likely sum to 1) could be based on a combination of the following metrics, assuming we are comparing two templates that are identical aside for one being methylated and the other being unmethylated at a given position: the ratio of or difference in IPD between the two templates at that given position; the statistical significance of the distinguishability between the IPD distributions of the two templates at that given positions; the number of observations used when creating the IPD distributions; and the neighboring sequence context. An example w_i could be $w_i = \log\left[\frac{\langle IPD_i \rangle_{Me+}}{\langle IPD_i \rangle_{Me-}}\right]$. This signal can also be weighted by the prior probability of seeing a Me+ signal.

[00147] In certain aspects, the invention provides a general-purpose approach to discriminating between Me+ and Me- using features in a real-time sequencing-by-synthesis trace comprising signals emitted during the incorporation of optically detectable nucleotides into a nascent strand by a polymerase enzyme. Such traces and various methods of analysis thereof are further described elsewhere, e.g., in U.S. Patent Publication No. 20090024331, incorporated

herein by reference in its entirety for all purposes. A first stage of this approach includes the development of a classifier for distinguishing methylated from unmethylated cytosines in a nucleic acid template. A set of features is measured for every pulse (discrete event or signal) in the trace. For example, a set of measurable features might be {pulse width, pulse duration, pulse height, pulse amplitude variability}. Call the values of these features $f_1^i, f_2^i, \ldots, f_k^i$ for pulse *i*. For each cytosine pulse in the methylated and unmethylated template data sets (which may or may not be restricted to CpG), tabulation is performed for the local pulse features $\vec{f} = \{f_1^{i-3}, f_2^{i-3}, \ldots, f_k^{i-1}, f_1^i, f_2^i, \ldots, f_k^i, \ldots, f_1^{i+3}, f_2^{i+3}, \ldots, f_k^{i+3}\}$. In this example, a local context extending 3 pulses to the left and right of the pulse of interest is assumed, but this context size is flexible and, in certain embodiments, can be application-specific.

[00148] The observed data likelihoods $p(\vec{f}|Me+)$ and $p(\vec{f}|Me-)$ are derived, e.g., by a kernel density method or simple binning and tabulation of the features. Thus, a generalized signal for determining methylation status on the trace has been determined: $w_i = \log\left[\frac{p(\vec{f}|Me+)p(Me+)}{p(\vec{f}|Me-)p(Me-)}\right]$, where $p(Me+)$ and $p(Me-)$ are the prior probabilities of methylated or un-methylated positions, respectively.

[00149] Various standard classification algorithm development techniques known to those of ordinary skill in the art may be applied to refine this approach, both to reduce training set bias and to improve sensitivity. Such techniques include but are not limited to cross-validation, boosting, and bootstrap aggregating (bagging). In certain embodiments, the set of feature inputs is restricted to those that are most correlated with the Me+ and Me- status of a position. In certain embodiments, the major component in a principal components analysis can serve as a better weighted combination of the most important features. In further embodiments, leave-one-out cross-validation can be valuable in selecting a robust predictive algorithm, e.g., by mitigating overfitting to the observed data that can occur when developing a classifier on a training set. Further, in some embodiments a boosting approach (training of a hierarchy of classifiers on the progressively more difficult regions of feature space) is applied to improve sensitivity.

[00150] More sophisticated signals can be employed to detect multiple, closely spaced CpGs. In certain embodiments, the data likelihoods described above can be measured for the case of two CpG sites with a known methylation state located a known distance apart, e.g., 2 base pairs apart. The signal generalizes to

$$w_i = \underset{\mu \in \{++,+-,-+,--\}}{\arg\max} \log\left[\frac{p(\vec{f}_\alpha, \vec{f}_\beta|\mu)p(\mu)}{\sum_{\mu'} p(\vec{f}_\alpha, \vec{f}_\beta|\mu')p(\mu')}\right],$$

where $p(\mu)$ and $p(\mu')$ are the prior probabilities of (methylated,un-methylated) configurations (this joint distribution would be assumed to be independent unless otherwise shown).

[00151] Although described primarily in the context of detection of methyl cytosine, these methods are also applicable to methyl-adenosine or any other base modification for which IPDs are used as a metric for detection. Figure 14 provides data showing differences between ratios of IPDs for methylated adenosines and unmethylated adenosines in a template nucleic acid, and the data used to generate Figure 14 is further described in the Examples herein. This data also shows that N^6-methyladenosine, like methylcytosine, has an effect on IPD not only at the methylated base but also at multiple, neighboring positions, as well. Further, in light of the above teachings it will be clear to one of ordinary skill that the approach can be extended to pulse metrics other than IPD, such as pulse width, branch rate, mismatch rate, deletion rate, etc. In addition, the general classifier approach suggested in steps 2+3 can be implemented with many standard statistical classification algorithms, i.e. linear discriminant analysis, multi-dimensional regression, kernel methods, classification and regression trees, neural networks, and support vector machines. The approach can also incorporate data from multiple strands of a duplex template. For example, because the CG sequence for cytosine methylation and the GATC sequence for adenosine methylation is the same on the reverse complement strand, these bases can be methylated on both complementary strands. If the general statistical distribution for the fraction of sites that are hemi-methylated vs. fully methylated is known, then information regarding IPD or other metrics gained from the complementary strand can be used to increase the accuracy with which a call is made on a particular strand. For example, if after analyzing each strand separately it is concluded that there is a 95% chance that stand A is methylated and a 55% chance that complementary strand B is methylated, but it is known that there is a 80% chance that if one strand is methylated then so is the other, then the confidence in calling strand B as methylated is increased.

[00152] Another modified base for which IPDs may be used as a metric for detection is 5-hydroxymethylcytosine (5-hmC). It was recently found to be abundant in human and mouse brains, as well as in embryonic stem cells (see, e.g., Kriaucionis, et al. (2009) "The nuclear DNA base 5-hydroxymethylcytosine is present in Purkinje neurons and the brain" *Science* **324** (5929): 929–30; and Tahiliani M et al. (May 2009) "Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1" *Science* **324** (5929): 930–

35, incorporated herein by reference in their entireties for all purposes). In mammals, it can be generated by oxidation of 5-methylcytosine, a reaction mediated by the Tet family of enzymes. Conventional bisulfite sequencing does not effectively distinguish 5-hmC from 5-MeC because 5-hmC tends to remain unmodified like 5-MeC. As such, mass spectrometry is the typical means of detecting 5-hmC in a nucleic acid sample. The methods described herein provide a high-throughput, real-time method to distinguish between C, 5-MeC, and 5-hmC by monitoring deviations from normal polymerase kinetics, including IPD and pulse width.

[00153] Experiments were carried out to test the ability of the methods of the invention to distinguish between 5-MeC and 5-hmC, and it was found that 5-hmC causes an increase in IPD at certain positions surrounding the 5-hmC site in the template, and also decreases the pulse width at that position. Further, the data generated suggests that 5-hmC may also increase the pulse width at the position following the 5-hmC site. Both the difference in IPD and the difference in pulse width between C and 5-hmC were larger in magnitude than were the differences in IPD and pulse width between C and 5-MeC, and these larger magnitudes are likely to make 5-hmC even more detectable than 5-MeC. Without being bound by theory, the reason for the higher magnitude differences for these two measures may be due to the additional oxygen atom present in 5-hmC as compared to 5-MeC. This additional oxygen could yield additional steric and charge-based interactions between the polymerase and the DNA template that slow the binding and/or incorporation of the complementary base into the nascent strand.

[00154] Based on the findings that indicate the easier detection of 5-hmC as compared to 5-MeC, in certain embodiments template nucleic acids can be treated to convert 5-MeC to the more easily detected 5-hmC, e.g., by treatment with an enzyme such as TET1, which converts 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA (see, e.g., Tahiliani M et al., *supra*). Although this technique would not permit distinction between 5-MeC and 5-hmC in the template (since the 5-MeC converted to 5-hmC would be indistinguishable from any 5-hmC originally present in the template), it will nonetheless be useful for facilitating detection 5-MeC patterns in template nucleic acids with the caveat that the patterns so discovered may, in vivo, also include 5-hmC bases.

[00155] In order to maximally use the IPD and pulse width signals from multiple positions surrounded the 5-hmC site, one could use a technique to find the optimal weighting of different positions for IPD and pulse width in order to distinguish 5-hmC, 5-MeC, and C from one another. An example of one such technique is principle component analysis, and others are known in the art. Principle component analysis can be described as finding the eigenvector (using each metric such as IPD or pulse width at each position as a different basis vector, such that if you have 10 positions in question and two metrics, your basis will have 2x10=20

dimensions) with the greatest eigenvalue. For a review of principle component analysis, see e.g. Jolliffe I.T. Principal Component Analysis, Series: Springer Series in Statistics, 2nd ed., Springer, NY, 2002, the disclosure of which is incorporated herein by reference in its entirety for all purposes.

IX. Systems

[00156] The invention also provides systems that are used in conjunction with the compositions and methods of the invention in order to provide for real-time single-molecule detection of analytical reactions. In particular, such systems typically include the reagent systems described herein, in conjunction with an analytical system, e.g., for detecting data from those reagent systems. In certain preferred embodiments, analytical reactions are monitored using an optical system capable of detecting and/or monitoring interactions between reactants at the single-molecule level. For example, such an optical system can achieve these functions by first generating and transmitting an incident wavelength to the reactants, followed by collecting and analyzing the optical signals from the reactants. Such systems typically employ an optical train that directs signals from the reactions to a detector, and in certain embodiments in which a plurality of reactions is disposed on a solid surface, such systems typically direct signals from the solid surface (e.g., array of confinements) onto different locations of an array-based detector to simultaneously detect multiple different optical signals from each of multiple different reactions. In particular, the optical trains typically include optical gratings or wedge prisms to simultaneously direct and separate signals having differing spectral characteristics from each confinement in an array to different locations on an array based detector, e.g., a CCD, and may also comprise additional optical transmission elements and optical reflection elements.

[00157] An optical system applicable for use with the present invention preferably comprises at least an excitation source and a photon detector. The excitation source generates and transmits incident light used to optically excite the reactants in the reaction. Depending on the intended application, the source of the incident light can be a laser, laser diode, a light-emitting diode (LED), a ultra-violet light bulb, and/or a white light source. Further, the excitation light may be evanescent light, e.g., as in total internal reflection microscopy, certain types of waveguides that carry light to a reaction site (see, e.g., U.S. Application Pub. Nos. 20080128627, 20080152281, and 200801552280), or zero mode waveguides, described below. Where desired, more than one source can be employed simultaneously. The use of multiple sources is particularly desirable in applications that employ multiple different reagent compounds having differing excitation spectra, consequently allowing detection of more than one fluorescent signal to track the interactions of more than one or one type of molecules simultaneously (e.g., multiple types of

differentially labeled reaction components). A wide variety of photon detectors or detector arrays are available in the art. Representative detectors include but are not limited to an optical reader, a high-efficiency photon detection system, a photodiode (e.g. avalanche photo diodes (APD)), a camera, a charge-coupled device (CCD), an electron-multiplying charge-coupled device (EMCCD), an intensified charge coupled device (ICCD), and a confocal microscope equipped with any of the foregoing detectors. For example, in some embodiments an optical train includes a fluorescence microscope capable of resolving fluorescent signals from individual sequencing complexes. Where desired, the subject arrays of optical confinements contain various alignment aides or keys to facilitate a proper spatial placement of the optical confinement and the excitation sources, the photon detectors, or the optical train as described below.

[00158] The subject optical system may also include an optical train whose function can be manifold and may comprise one or more optical transmission or reflection elements. Such optical trains preferably encompass a variety of optical devices that channel light from one location to another in either an altered or unaltered state. First, the optical train collects and/or directs the incident wavelength to the reaction site (e.g., optical confinement). Second, it transmits and/or directs the optical signals emitted from the reactants to the photon detector. Third, it may select and/or modify the optical properties of the incident wavelengths or the emitted wavelengths from the reactants. Illustrative examples of such optical transmission or reflection elements are diffraction gratings, arrayed waveguide gratings (AWG), optical fibers, optical switches, mirrors (including dichroic mirrors), lenses (including microlenses, nanolenses, objective lenses, imaging lenses, and the like), collimators, optical attenuators, filters (e.g., polarization or dichroic filters), prisms, wavelength filters (low-pass, band-pass, or high-pass), planar waveguides, wave-plates, delay lines, and any other devices that guide the transmission of light through proper refractive indices and geometries. One example of a particularly preferred optical train is described in U.S. Patent Pub. No. 20070036511, filed August 11, 2005, and incorporated by reference herein in its entirety for all purposes.

[00159] In a preferred embodiment, a reaction site (e.g., optical confinement) containing a reaction of interest is operatively coupled to a photon detector. The reaction site and the respective detector can be spatially aligned (e.g., 1:1 mapping) to permit an efficient collection of optical signals from the reactants. In certain preferred embodiments, a reaction substrate is disposed upon a translation stage, which is typically coupled to appropriate robotics to provide lateral translation of the substrate in two dimensions over a fixed optical train. Alternative embodiments could couple the translation system to the optical train to move that aspect of the system relative to the substrate. For example, a translation stage provides a means of removing a

reaction substrate (or a portion thereof) out of the path of illumination to create a non-illuminated period for the reaction substrate (or a portion thereof), and returning the substrate at a later time to initiate a subsequent illuminated period. An exemplary embodiment is provided in U.S. Patent Pub. No. 20070161017, filed December 1, 2006.

[00160] In particularly preferred aspects, such systems include arrays of reaction regions, e.g., zero mode waveguide arrays, that are illuminated by the system, in order to detect signals (e.g., fluorescent signals) therefrom, that are in conjunction with analytical reactions being carried out within each reaction region. Each individual reaction region can be operatively coupled to a respective microlens or a nanolens, preferably spatially aligned to optimize the signal collection efficiency. Alternatively, a combination of an objective lens, a spectral filter set or prism for resolving signals of different wavelengths, and an imaging lens can be used in an optical train, to direct optical signals from each confinement to an array detector, e.g., a CCD, and concurrently separate signals from each different confinement into multiple constituent signal elements, e.g., different wavelength spectra, that correspond to different reaction events occurring within each confinement. In preferred embodiments, the setup further comprises means to control illumination of each confinement, and such means may be a feature of the optical system or may be found elsewhere is the system, e.g., as a mask positioned over an array of confinements. Detailed descriptions of such optical systems are provided, e.g., in U.S. Patent Pub. No. 20060063264, filed September 16, 2005, which is incorporated herein by reference in its entirety for all purposes.

[00161] The systems of the invention also typically include information processors or computers operably coupled to the detection portions of the systems, in order to store the signal data obtained from the detector(s) on a computer readable medium, e.g., hard disk, CD, DVD or other optical medium, flash memory device, or the like. For purposes of this aspect of the invention, such operable connection provides for the electronic transfer of data from the detection system to the processor for subsequent analysis and conversion. Operable connections may be accomplished through any of a variety of well known computer networking or connecting methods, e.g., Firewire®, USB connections, wireless connections, WAN or LAN connections, or other connections that preferably include high data transfer rates. The computers also typically include software that analyzes the raw signal data, identifies signal pulses that are likely associated with incorporation events, and identifies bases incorporated during the sequencing reaction, in order to convert or transform the raw signal data into user interpretable sequence data (see, e.g., Published U.S. Patent Application No. 2009-0024331, the full disclosure of which is incorporated herein by reference in its entirety for all purposes).

[00162] Exemplary systems are described in detail in, e.g., U.S. Patent Application No. 11/901,273, filed September 14, 2007 and U.S. Patent Application No. 12/134,186, filed June 5, 2008, the full disclosures of which are incorporated herein by reference in their entirety for all purposes.

[00163] Further, the invention provides data processing systems for transforming raw data generated in an analytical reaction into analytical data that provides a measure of one or more aspects of the reaction under investigation, e.g., transforming signals from a sequencing-by-synthesis reaction into nucleic acid sequence read data, which can then be transformed into consensus sequence data. In certain embodiments, the data processing systems include machines for generating nucleic acid sequence read data by polymerase-mediated processing of a template nucleic acid molecule (e.g., DNA or RNA). The nucleic acid sequence read data generated is representative of the nucleic acid sequence of the nascent polynucleotide synthesized by a polymerase translocating along a nucleic acid template only to the extent that a given sequencing technology is able to generate such data, and so may not be identical to the actual sequence of the nascent polynucleotide molecule. For example, it may contain a deletion or a different nucleotide at a given position as compared to the actual sequence of the polynucleotide, e.g., when a nucleotide incorporation is missed or incorrectly determined, respectively. As such, it is beneficial to generate redundant nucleic acid sequence read data, and to transform the redundant nucleic acid sequence read data into consensus nucleic acid sequence data that is generally more representative of the actual sequence of the polynucleotide molecule than nucleic acid sequence read data from a single read of the nucleic acid molecule. Redundant nucleic acid sequence read data comprises multiple reads, each of which includes at least a portion of nucleic acid sequence read that overlaps with at least a portion of at least one other of the multiple nucleic acid sequence reads. As such, the multiple reads need not all overlap with one another, and a first subset may overlap for a different portion of the nucleic acid sequence than does a second subset. Such redundant sequence read data can be generated by various methods, including repeated synthesis of nascent polynucleotides from a single nucleic acid template, synthesis of polynucleotides from multiple identical nucleic acid templates, or a combination thereof.

[00164] In another aspect, the data processing systems can include software and algorithm implementations provided herein, e.g. those configured to transform redundant nucleic acid sequence read data into consensus nucleic acid sequence data, which, as noted above, is generally more representative of the actual sequence of the nascent polynucleotide molecule than nucleic acid sequence read data from a single read of a single nucleic acid molecule. Further, the transformation of the redundant nucleic acid sequence read data into consensus nucleic acid sequence data identifies and negates some or all of the single-read variation

between the multiple reads in the redundant nucleic acid sequence read data. As such, the transformation provides a representation of the actual nucleic acid sequence of the nascent polynucleotide complementary to the nucleic acid template that is more accurate than a representation based on a single read.

[00165] Various methods and algorithms for data transformation employ data analysis techniques that are familiar in a number of technical fields, and are generally referred to herein as statistical analysis. For clarity of description, details of known techniques are not provided herein. These techniques are discussed in a number of available reference works, such as those provided in U.S. Patent Publication No. 20090024331 and U.S.S.N. 61/116,439, filed November 20, 2008, the disclosures of which are incorporated herein by reference in their entireties for all purposes.

[00166] The software and algorithm implementations provided herein are preferably machine-implemented methods, e.g., carried out on a machine comprising computer-readable medium configured to carry out various aspects of the methods herein. For example, the computer-readable medium preferably comprises at least one or more of the following: a) a user interface; b) memory for storing raw analytical reaction data; c) memory storing software-implemented instructions for carrying out the algorithms for transforming the raw analytical reaction data into transformed data that characterizes one or more aspects of the reaction (e.g., rate, consensus sequence data, etc.); d) a processor for executing the instructions; e) software for recording the results of the transformation into memory; and f) memory for recordation and storage of the transformed data. In preferred embodiments, the user interface is used by the practitioner to manage various aspects of the machine, e.g., to direct the machine to carry out the various steps in the transformation of raw data into transformed data, recordation of the results of the transformation, and management of the transformed data stored in memory.

[00167] As such, in preferred embodiments, the methods further comprise a transformation of the computer-readable medium by recordation of the raw analytical reaction data and/or the transformed data generated by the methods. Further, the computer-readable medium may comprise software for providing a graphical representation of the raw analytical reaction data and/or the transformed data, and the graphical representation may be provided, e.g., in soft-copy (e.g., on an electronic display) and/or hard-copy (e.g., on a print-out) form.

[00168] The invention also provides a computer program product comprising a computer-readable medium having a computer-readable program code embodied therein, the computer readable program code adapted to implement one or more of the methods described herein, and optionally also providing storage for the results of the methods of the invention. In certain

preferred embodiments, the computer program product comprises the computer-readable medium described above.

[00169] In another aspect, the invention provides data processing systems for transforming raw analytical reaction data from one or more analytical reactions into transformed data representative of a particular characteristic of an analytical reaction, e.g., an actual sequence of one or more template nucleic acids analyzed, a rate of an enzyme-mediated reaction, an identity of a kinase target molecule, and the like. Such data processing systems typically comprise a computer processor for processing the raw data according to the steps and methods described herein, and computer usable medium for storage of the raw data and/or the results of one or more steps of the transformation, such as the computer-readable medium described above.

[00170] As shown in Figure 9, the system 900 includes a substrate 902 that includes a plurality of discrete sources of chromophore emission signals, e.g., an array of zero mode waveguides 904. An excitation illumination source, e.g., laser 906, is provided in the system and is positioned to direct excitation radiation at the various signal sources. This is typically done by directing excitation radiation at or through appropriate optical components, e.g., dichroic 908 and objective lens 910, that direct the excitation radiation at the substrate 902, and particularly the signal sources 904. Emitted signals from the sources 904 are then collected by the optical components, e.g., objective 910, and passed through additional optical elements, e.g., dichroic 908, prism 912 and lens 914, until they are directed to and impinge upon an optical detection system, e.g., detector array 916. The signals are then detected by detector array 916, and the data from that detection is transmitted to an appropriate data processing system, e.g., computer 918, where the data is subjected to interpretation, analysis, and ultimately presented in a user ready format, e.g., on display 920, or printout 922, from printer 924. As will be appreciated, a variety of modifications may be made to such systems, including, for example, the use of multiplexing components to direct multiple discrete beams at different locations on the substrate, the use of spatial filter components, such as confocal masks, to filter out-of focus components, beam shaping elements to modify the spot configuration incident upon the substrates, and the like (See, e.g., Published U.S. Patent Application Nos. 2007/0036511 and 2007/095119, and U.S. Patent Application No. 11/901,273, all of which are incorporated herein by reference in their entireties for all purposes.)

II. Examples

Detection of 5-methylcytosine (5-MeC)

[00171] Methylation sequencing on a SMRT™ Sequencing platform (see, e.g., P. M. Lundquist, et al., *supra*) was performed on short, synthetic DNA oligos with contrived patterns

of methylated and unmethlyated bases, along with control sequences having the same primary sequence but without any methylation. These templates provided unequivocal fluorescence pulse patterns and tempos that demonstrated how the combination of sequence context and methylation status affected interpulse duration. For example, SMRT™ sequencing experiments were performed using synthetic DNA templates that only differed by a single methylated *vs.* unmethylated cytosine. The difference in average interpulse durations between the two templates was visible both at the 5-MeC position and in the vicinity of the 5-MeC position.

[00172] Because the interpulse duration between any two successive incorporation events is stochastic in nature and has an exponential distribution (Eid, et al., *supra*), a single sequencing measurement may not always yield enough information to determine methylation status with certainty. Therefore, in certain embodiments a highly processive, strand-displacing polymerase is used, and this polymerase carries out multiple laps of synthesis around a circular DNA template (J. Korlach, et al., *Proc Natl Acad Sci U S A* **2008**, *supra*). This mode of operation provides repeated sequencing of the same DNA molecule to generate multiple sequence reads, e.g., by rolling circle replication. The statistical distribution of interpulse durations obtained at a particular template site will thus indicate its methylation state.

[00173] In particular, Figure 10A shows a schematic of two templates for use in SMRT™ sequencing. Both are comprises a double-stranded region flanked by two single-stranded hairpins. A polymerase binds to a primed location on the template, e.g., via a primer hybridized to one of the single-stranded hairpins, and commences processing the template to generate a nascent strand complementary to the strand upon which the polymerase is translocating. The strand displacement activity of the polymerase permits passage through the double-stranded region which is unwound to transform the template into a circular form. The polymerase then proceeds around the other single-stranded hairpin and on through the previously displaced strand of the double-stranded region. The polymerase can continue to process the template in a "rolling-circle" fashion to generate a concatemer comprising multiple copies of complements to both strands of the double-stranded region, as well as the hairpins. The two templates are identical except at position 2, where the top template comprises a methylated cytosine (5-MeC) and the bottom template comprises a non-methylated cytosine. (Position 1 is a non-methylated cytosine in both templates.) Figure 10B provides an illustrative depiction of the difference in IPD for the methylated template as compared to the unmethylated template. For each row, the histograms depict the distributions of mean weighted IPD (averaged over the labeled number of circular consensus sequencing subreads (in this context, a sequence read generated from a single pass of the polymerase around the template). Specifically, "1" indicates the sequencing data was derived from a sequence read generated in a single pass around the template; "3" indicates the

data was derived from a sequence read generated in three passes around the template; and "5" indicates the data was derived from a sequence read generated in five passes around the template. The data from the methylated template is shown as a solid line, and the data from the unmethylated template is shown as a dotted line. At Position 1, the distributions of weighted IPD for the two templates are very similar. At Position 2, the average weighted IPD after a single subread (top histogram) is longer in the methylated template than in the unmethylated template. After 3 and 5 circular subreads, the distributions overlap even less. The interpulse duration (IPD) was clearly lengthened by the presence of 5-MeC. These results demonstrated the ability to use SMRT™ sequencing technology to perform methylation sequencing of DNA. Weighted IPDs are described elsewhere herein.

[00174] Further, methylcytosine was shown to have an effect on interpulse duration (IPD) not only at the methylated base, but over a range of several bases upstream and downstream of the position of the methylcytosine. Specifically, an increase in IPD was observed at some positions in the presence of methylcytosine relative to the same position in the absence of methylcytosine. Figure 11 provides a plot depicting the ratio of the average IPD in the methylated template to the average IPD in the unmethylated template, plotted versus DNA template position. The two templates are identical except for the methylated bases in the methylated template, which are indicated by arrowheads in Figure 11.

[00175] Figure 12 provides another data set illustrating the ratio of IPD for a different methylated template vs. an identical but unmethylated template as a function of position. Seven cytosines (shown with crosshatching) were differentially methylated (5-MeC) between the two templates. That data clearly showed that IPD was increased in the region comprising the methylated bases. Interestingly, the effect on IPD occurred mostly downstream of the methylated positions. As such, data from nascent strand synthesis at positions in the template that are near the differentially methylated site, in addition to the differentially methylated site itself, is useful for methylation detection during real-time nascent strand synthesis.

Detection of N6-methyladenosine (N6-MeA)

[00176] Similar methods as those used to detect 5-MeC were used to detect N6-MeA in similarly constructed template nucleic acids. Figure 13A shows a schematic of two templates, both of which comprise a double-stranded region flanked by two single-stranded hairpins. The methylated template has an A within a GATC context at Position 1 and a ^{m}A within a GATC context at Position 2, whereas the unmethylated template has an A at both positions. Otherwise, the two templates are identical. As described above, a polymerase binds to a primed location on the template and commences processing the template to generate a nascent strand, using its strand displacement activity to unwind the double-stranded region and proceed around the

template. Figure 13B shows plots of mean IPD generated from sequencing data using these two templates for varying numbers of consensus reads, as described above. The data from the methylated template is shown as a solid line, and the data from the unmethylated template is shown as a dotted line. For each row in Figure 13B, the histograms depict the distributions of mean IPD (averaged over the labeled number of consensus sequencing subreads, i.e. the number of times the polymerase made one complete pass around the template to generate a complementary nascent strand). At Position 1, the distributions of IPD for the two templates are very similar. At Position 2, the average IPD after a single subread (top histogram) is ~5× longer in the methylated template than in the unmethylated template. After 3 and 5 circular subreads, the distributions overlap even less. The interpulse duration (IPD) was clearly lengthened by the presence of N6-MeA, demonstrating that the SMRT™ sequencing technology can be used to perform methylation sequencing of DNA comprising a methylated base other than 5-MeC.

[00177] Receiver operating characteristic (ROC) curves, parameterized by IPD threshold, for assigning a methylation status to an adenosine nucleotide are provided in Figure 13C. True positive means that an ^{m}A is correctly called as ^{m}A, whereas a false positive means that an A is mistakenly called as ^{m}A. These ROC curves, based on the IPD distributions from Position 2 in Figure 13B, are shown for a single read (solid line), and for 3 (long-dashed line) or 5 (short-dashed line) molecular redundant sequencing reads produced by the polymerase processing the template one, three, or five times, respectively. The dotted horizonal line bisecting the graph depicts the ROC curve for randomly guessing the methylation status. The normalized area under the ROC curve is 0.80 after the first circular subread but increases to 0.92 and 0.96 after three and five circular subreads, respectively. In fact, after five subreads, >85% of ^{m}A bases can be detected at this template position with a false positive rate of only ~5%.

[00178] Like methylcytosine, methyladenosine was also shown to have an effect on IPD over a range of several bases upstream and downstream of the position of the methyladenosine. Specifically, an increase in IPD was observed at some positions in the presence of methyladenosine relative to the same position in the absence of methyladenosine. Figure 14 provides a plot depicting the ratio of the average IPD in the methylated template to the average IPD in the unmethylated template, plotted versus DNA template position. The two templates are identical except for the methylated bases in the methylated template, which are indicated by arrowheads in Figure 14.

Detection of 5-hydroxymethylcytosine

[00179] Similar to 5-MeC and N6-MeA, 5-hydroxymethylcytosine was also tested and shown to have an effect on IPD over a range of several bases upstream and downstream of the position of the 5-hydroxymethylcytosine. Specifically, an increase in IPD was observed at some

positions in the presence of 5-hydroxymethylcytosine relative to the same position in the absence of 5-hydroxymethylcytosine. Figure 15 provides a plot depicting the ratio of the average IPD in the hydroxymethylated template to the average IPD in the unmethylated template, plotted versus DNA template position. The two templates are identical except for the hydroxymethylcytosine in the hydroxymethylated template, which are indicated by arrowheads in Figure 15. Templates comprising 5-hydroxymethylcytosine bases were also tested and the presence of these modifications was shown to have an effect on pulse width. Figure 16 provides a plot of pulse width ratio (pulse width for methylated template divided by pulse width for unmethylated template) vs. template position where the modified positions comprise 5-hydroxymethylcytosine bases. Variably hydroxymethylated positions are indicated by the arrowheads.

Detection of 8-oxoguanosine (8-oxoG)

[00180] 8-oxoguanosine was also subjected to single molecule real-time sequencing and was shown to affect IPD both at the site of the modifications as well as at proximal unmodified positions in the template. An increase in IPD was observed at some positions in the presence of 8-oxoguanosine relative to the same position in the absence of 8-oxoguanosine. Figure 17 provides a plot depicting the ratio of the average IPD in the 8-oxoguanosine template to the average IPD in the unmodified template, plotted versus DNA template position. The two templates are identical except for the 8-oxoguanosine in the modified template, which are indicated by arrowheads in Figure 17. These data showed that, compared to G, 8-oxoG altered IPD significantly over a window of ~10 neighboring bases surrounding the 8-oxoG position. Some positions saw an increase in IPD by a factor of as much as 6.5x. Templates comprising 8-oxoG bases were also tested and the presence of these modifications was shown to have an effect on pulse width. Figure 18 provides a plot of pulse width ratio (pulse width for an 8-oxoG template divided by pulse width for template with no 8-oxoG) vs. template position where the modified positions comprise 8-oxoG bases. Variable positions are indicated by the arrowheads. Further, 8-oxoG altered pulse width over a window of 7-8 neighboring bases by as much as 40%, and such alteration included both increases and decreases in pulse width.

[00181] It is to be understood that the above description is intended to be illustrative and not restrictive. It readily should be apparent to one skilled in the art that various embodiments and-modifications may be made to the invention disclosed in this application without departing from the scope and spirit of the invention. The scope of the invention should, therefore, be determined not with reference to the above description, but should instead be determined with reference to the appended claims, along with the full scope of equivalents to which such claims are entitled. All publications mentioned herein are cited for the purpose of describing and

disclosing reagents, methodologies and concepts that may be used in connection with the present invention. Nothing herein is to be construed as an admission that these references are prior art in relation to the inventions described herein. Throughout the disclosure various patents, patent applications, and publications are referenced. 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.

What is claimed is:

1. A method for identifying a modification, the method comprising:
   a) providing a template nucleic acid comprising the modification;
   b) providing an enzyme capable of processing the template nucleic acid;
   c) contacting the template nucleic acid with the enzyme;
   d) monitoring processing of the template nucleic acid by the enzyme; and
   e) detecting a change in the processing, wherein the change is indicative of the modification, thereby indentifying the modification.

2. The method of claim 1, wherein the modification is selected from the group consisting of a methylated base, a pseudouridine base, an 7,8-dihydro-8-oxoguanine base, a 2'-O-methyl derivative base, a nick, an apurinic site, an apyrimidic site, a pyrimidine dimer, a cis-platen crosslinking, oxidation damage, hydrolysis damage, a bulky base adduct, a thymine dimer, a photochemistry reaction product, an interstrand crosslinking product, a mismatched base, a secondary structure, or a bound agent.

3. The method of claim 2, wherein the modification is a methylated cytosine or adenine base.

4. The method of claim 1, wherein the template nucleic acid comprises a single-stranded portion and a double-stranded portion.

5. The method of claim 4, wherein the double-stranded portion is a result of complementarity between two separate portions of the template nucleic acid.

6. The method of claim 1, wherein the template nucleic acid comprises a first polynucleotide region comprising the modification and a second polynucleotide region complementary to the first polynucleotide region, where the first polynucleotide region and the second polynucleotide region are on a single strand of the template nucleic acid.

7. The method of claim 1, wherein the template nucleic acid is subjected to a treatment to alter the modification prior to the contacting,

8. The method of claim 7, wherein the treatment is selected from the group consisting of: glycosylase modification, bisulfite modification, DMS modification, cytosine methyltransferase modification, TET1 modification, restriction enzyme modification, and cytidine deaminase modification.

9. The method of claim 8, wherein the treatment is glycosylase modification to convert the modification to an abasic site, and further wherein the processing results in the synthesis of a

nascent nucleic acid strand into which a non-natural nucleotide analog is incorporated opposite the abasic site.

10. The method of claim 9, wherein the non-natural nucleotide analog comprises a detectable label that distinguishes its incorporation from incorporation of other nucleotides into the nascent nucleic acid strand.

11. The method of claim 9, wherein the non-natural nucleotide analog is a pyrene analog.

12. The method of claim 9, wherein the non-natural nucleotide analog is a mixture of incorporatable non-natural nucleotide analogs that are non-incorporatable non-natural nucleotide analogs.

13. The method of claim 12, wherein the non-incorporatable non-natural nucleotide analogs associates with the abasic site multiple times prior to incorporation of one of the incorporatable non-natural nucleotide analogs.

14. The method of claim 8, wherein the treatment is DMS modification and the modification is secondary structure of the template nucleic acid or an agent bound to the template nucleic acid.

15. The method of claim 1, wherein the template nucleic acid is a circular nucleic acid.

16. The method of claim 15, wherein the processing is rolling-circle synthesis of a nascent nucleic acid strand.

17. The method of claim 1, wherein the template nucleic acid is an RNA or DNA molecule.

18. The method of claim 1, wherein the enzyme is a polymerase enzyme.

19. The method of claim 7, wherein the polymerase enzyme is selected from the group consisting of a DNA polymerase, an RNA polymerase, a reverse transcriptase, and a derivative thereof.

20. The method of claim 7, wherein the polymerase enzyme is a Φ29 polymerase comprising at least one mutation at a position selected from the group consisting of K392, K422, I93, M188, K392, V399, T421, K422; S95, Y101, M102; Q99, L123, K124, T189, A190; G191, S388; P127, L384, N387, S388; and L389, Y390, and G391.

21. The method of claim 7, wherein the polymerase enzyme displays strand displacement activity.

22. The method of claim 1, wherein the processing results in the synthesis of a nascent nucleic acid strand.

23. The method of claim 1, wherein the processing is a single-molecule sequencing reaction.

24. The method of claim 22, wherein the monitoring detects incorporation of single nucleotides into the nascent nucleic acid strand to generate a sequence read that is complementary to the template nucleic acid.

25. The method of claim 24, wherein the change occurs at the modification.

26. The method of claim 24, wherein the change occurs at one or more positions upstream or downstream of the modification.

27. The method of claim 24, wherein the change comprises incorporation of a non-natural nucleotide analog at a position in the nascent nucleic acid strand that is complementary to a position in the template nucleic acid that comprises the modification.

28. The method of claim 27, wherein the change further comprises a kinetic difference in the processing.

29. The method of claim 27, wherein the modification is a methylated cytosine and the non-natural nucleotide analog is a modified guanine nucleotide analog.

30. The method of claim 27, wherein the modification is a modified base that forms a non-natural base pair with the non-natural nucleotide analog, the non-natural base pair being selected from the group consisting of isocytosine and isoguanine; 5-methylisocytosine and isoguanine; Im-N^O and Im-O^N; A* and T*; and 8-oxoG and adenine.

31. The method of claim 27, wherein the non-natural nucleotide analog comprises a detectable label that distinguishes its incorporation from incorporation of other nucleotides into the nascent nucleic acid strand.

32. The method of claim 24, wherein the single nucleotides are differentially labeled to be distinguishable from one another during the monitoring.

33. The method of claim 32, wherein the single nucleotides comprise a label linked to a phosphate group, wherein the phosphate group is removed during the processing.

34. The method of claim 24, wherein the sequence read contains multiple copies of at least a portion complementary to the template nucleic acid.

35. The method of claim 24, further comprising mapping the modification within the template nucleic acid, the mapping comprising:

f) analyzing a portion of the sequence read that was generated immediately prior to, during, or immediately after the detecting the change in processing to determine a sequence complementary to the template nucleic acid;

g) determining the complement of the sequence complementary to the template nucleic acid in f; and

h) mapping the modification at a position in the template nucleic acid that is proximal to the complement of the sequence complementary to the template nucleic acid in f.

36. The method of claim 1, wherein the change is a kinetic difference in the processing.

37. The method of claim 36, wherein the kinetic difference is detected as an alteration in interpulse duration during the processing.

38. The method of claim 36, wherein the kinetic difference is detected as an alteration in pulse width during the processing.

39. The method of claim 36, wherein the kinetic difference is indicative of what type of modification is present in the template nucleic acid.

40. The method of claim 1, wherein the monitoring occurs in real time during the processing.

41. The method of claim 1, wherein the template nucleic acid and the enzyme form a complex that is immobilized at a reaction site on a substrate.

42. The method of claim 41, wherein a plurality of complexes are immobilized at optically resolvable reaction sites on the substrate, wherein a single complex immobilized at one of the reaction sites is optically resolvable from any other of the complexes immobilized at any other of the reaction sites.

43. The method of claim 42, wherein the optically resolvable reaction sites are nanometer-scale apertures in the substrate.

44. The method of claim 43, wherein the nanometer-scale apertures are optical confinements.

45. The method of claim 44, wherein the optical confinements are zero-mode waveguides.

46. The method of claim 1, wherein the template nucleic acid is plurality of template nucleic acids that are optically resolvable from one another during the monitoring.

47. The method of claim 1, wherein the template nucleic acid is not amplified prior to the contacting.

48. The method of claim 1, wherein the modification is a secondary structure formation and the change is an alteration in one or more kinetic characteristics of the processing, the method further comprising:

f) generating a sequence read for the template nucleic acid before and after the change;

g) identifying a first portion of the sequence read generated before the change that is complementary to a second portion of the sequence read generated after the change; and

h) determining a likelihood that a secondary structure formed by association of the first portion to the second portion was present in the template nucleic acid during the processing based at least upon the nucleotide composition of the first portion and the second portion.

49. A method of detecting binding of an agent to a single nucleic acid template, comprising:

a) providing the single nucleic acid template in complex with a polymerase;

b) introducing a reaction mixture to the complex, wherein the reaction mixture comprises the agent; and

c) monitoring synthesis of a polynucleotide by the polymerase, wherein the polynucleotide is complementary to the single nucleic acid template, and wherein a change in the synthesis is indicative of binding of the agent to the single nucleic acid template.

50. The method of claim 49, wherein the agent is selected from the group consisting of a transcription factor, a polymerase, a reverse transcriptase, a histone, a restriction enzyme, an antibody, a nucleic acid binding protein, and a nucleic acid binding agent.

51. The method of claim 49, wherein the single nucleic acid template is selected from the group consisting of a double-stranded DNA, a double-stranded RNA, a single-stranded DNA, a single-stranded RNA, a DNA/RNA hybrid, and a combination thereof.

52. The method of claim 49, further comprising determining a consensus binding site of the agent, said determining comprising:

d) performing a plurality of sequencing-by-synthesis reactions on a set of single nucleic acid templates in the presence of the agent to generate a set of binding-affected nascent polynucleotide sequences;

e) performing a plurality of sequencing-by-synthesis reactions on the set of single nucleic acid templates in the absence of the agent to generate a set of full-length nascent polynucleotide sequences;

f) analyzing the binding-affected nascent polynucleotide sequences to determine a location at which the agent bound the single nucleic acid template during the sequencing-by-synthesis reactions in the presence of the agent; and

g) identifying a sequence common to the full-length nascent polynucleotide sequences at the location determined in (iii), thereby identifying the consensus binding site of the agent.

53. The method of claim 52, wherein the binding-affected nascent polynucleotide sequences are truncated nascent polynucleotide sequences.

54. The method of claim 52, wherein the binding-affected nascent polynucleotide sequences are nascent polynucleotide sequences whose synthesis was paused at the location determined in f.

55. A method for detecting a modification in a single nucleic acid template during a sequencing-by-synthesis reaction, the method comprising:

a) providing the single nucleic acid template in complex with a polymerase;

b) introducing a reaction mixture to the complex, wherein the reaction mixture comprises an agent that specifically binds to the modification; and

c) monitoring synthesis of a polynucleotide by the polymerase, wherein the polynucleotide is complementary to the single nucleic acid template, and wherein a pause or cessation of the synthesis of the polynucleotide is indicative of binding of the agent to the single nucleic acid template, thereby detecting the modification in the single nucleic acid template.

56. The method of claim 55, wherein the modification is an 8-oxoG lesion.

57. The method of claim 55, wherein the agent is a protein is selected from the group consisting of hOGG1, FPG, yOGG1, AlkA, Nth, Nei, MutY, UDG, SMUG, TDG, NEIL, an antibody against 8-oxoG, or a binding domain thereof.

58. The method of claim 55, wherein the modification is a methylated base.

59. The method of claim 58, wherein the agent is a protein selected from the group consisting of MECP2, MBD1, MBD2, MBD4, UHRF1, an antibody against the methylated base, or a binding domain thereof.

60. The method of claim 55, wherein the modification is a secondary structure formation in the nucleic acid template.

61. The method of claim 55, wherein the complex is immobilized in an optical confinement.

62. The method of claim 55, wherein the template is selected from the group consisting of a single-stranded linear nucleic acid template, a single-stranded circular nucleic acid template, double-stranded linear nucleic acid template, a double-stranded circular nucleic acid template, and a combination thereof.

63. The method of claim 55, further comprising repairing the modification by including components of damage repair machinery in the reaction mixture.

64. The method of claim 63, wherein the readlength of the sequencing-by-synthesis reaction is longer than that for a further sequencing-by-synthesis reaction performed with the single nucleic acid template in complex with the polymerase, but absent the agent and the damage repair machinery.

65. A method for bypassing one or more modifications in a single nucleic acid template during a sequencing-by-synthesis reaction, the method comprising:

a) providing the single nucleic acid template in complex with a sequencing engine;

b) introducing a reaction mixture to the complex, wherein the reaction mixture comprises a bypass polymerase;

c) initiating the sequencing-by-synthesis reaction;

d) monitoring synthesis of a polynucleotide by the sequencing engine, wherein the polynucleotide is complementary to the single nucleic acid template, and wherein a pause or cessation of the synthesis of the polynucleotide is indicative that the sequencing engine has encountered a modification in the single nucleic acid template;

e) subsequently monitoring synthesis of the polynucleotide by the bypass polymerase, which is indicative that the modification is being bypassed; and

f) repeating steps c) and d) each time a further modification is encountered in the single nucleic acid template, thereby bypassing one or more modifications in a single nucleic acid template during a sequencing-by-synthesis reaction.

66. The method of claim 65, wherein the bypass polymerase comprises a detectable label and detection of a signal from the detectable label during the sequencing-by-synthesis reaction is indicative that the bypass polymerase is actively synthesizing the polynucleotide.

67. The method of claim 65, wherein the readlength of the sequencing-by-synthesis reaction is longer than that for a further sequencing-by-synthesis reaction performed with the single nucleic acid template in complex with the sequencing engine, but absent the bypass polymerase.

68. The method of claim 65, wherein the reaction mixture comprises multiple different bypass polymerases and a processivity factor.

69. The method of claim 65, wherein at least one of the single nucleic acid template, the sequencing engine, and the bypass polymerase is immobilized, directly or indirectly, in an optical confinement.

70. The method of claim 69, wherein the single nucleic acid template is immobilized by hybridization to an oligonucleotide primer immobilized in the optical confinement.

71. The method of claim 65, wherein the single nucleic acid template is processed by the sequencing engine multiple times at a single reaction site, and further wherein redundant sequence data is generated during the monitoring of step d).

72. The method of claim 71, wherein the single nucleic acid template is a circular nucleic acid template.

73. The method of claim 65, wherein the single nucleic acid template comprises multiple copies of a nucleic acid segment of interest, and the sequencing-by-synthesis reaction generates a polynucleotide comprising multiple copies of a segment complementary to the segment of interest, and further wherein redundant sequence data is generated during the monitoring of step d).

74. A composition comprising:

a) a substrate comprising a reaction site that is optically resolvable form any other reaction site on the substrate;

b) a single complex immobilized at the reaction site, wherein the complex comprises a template nucleic acid and a sequencing engine;

c) a mixture of incorporatable nucleotides or nucleotide analogs; and

d) at least one modification in the template nucleic acid, wherein the template at or proximal to the modification is processed differently than the template distal from the modification.

75. The composition of claim 74, wherein the modification is a non-natural base in the template nucleic acid.

76. The composition of claim 74, wherein the modification is in a strand of the template nucleic acid that is complementary to a nascent strand synthesized by the sequencing engine.

77. The composition of claim 74, wherein the modification is in a strand of the template nucleic acid that is displaced by the sequencing engine.

78. The composition of claim 74, further comprising at least one type of non-incorporatable nucleotide analog.

79. The composition of claim 74, wherein the mixture of incorporatable nucleotides or nucleotide analogs comprises at least one type of non-natural nucleotide analog.

80. The composition of claim 74, wherein the mixture of incorporatable nucleotides or nucleotide analogs comprises different types of nucleotides or nucleotide analogs, and further wherein the different types are distinguishable from one another because each type carries a distinct detectable label.

81. The composition of claim 74, further comprising an agent other than the sequencing engine that binds to the modification.

82. The composition of claim 74, further comprising an agent that chemically or enzymatically alters the modification.

83. The composition of claim 74, further comprising a nascent strand generated by the sequencing engine, wherein the nascent strand is complementary to the template nucleic acid.

84. The composition of claim 83, wherein the nascent strand comprises multiple copies of regions complementary to the template nucleic acid.

85. The composition of claim 74, further comprising a nanometer-scale aperture in the substrate, wherein the reaction site is disposed within the nanometer-scale aperture.

86. The composition of claim 74, wherein the template nucleic acid comprises internally complementary regions, and further wherein the modification is located within one of the internally complementary regions.

87. The composition of claim 85, wherein the nanometer-scale aperture is a ZMW.

88. A system for identification of modifications within a nucleic acid template, the method comprising:

a) a solid support having a polymerase complex disposed thereon, the polymerase complex comprising a nucleic acid template comprising a modification;
b) a mounting stage configured to receive the solid support;
c) an optical train positioned to be in optical communication with at least a portion of the solid support to detect signals emanating therefrom;
d) a translation system operably coupled to the mounting stage or the optical train for moving one of the optical train and the solid support relative to the other; and
e) a data processing system operably coupled to the optical train.

89. The system of claim 88, wherein the polymerase complex comprises a polymerase enzyme that is actively processing the nucleic acid template.

90. The system of claim 89, wherein the polymerase complex comprises a polymerase enzyme that is processively synthesizing a nascent strand by template-directed synthesis.

91. The system of claim 89, wherein the optical train detects signals emanating from the solid support during the processing of the nucleic acid template.

92. The system of claim 88, wherein the solid support comprises at least one zero-mode waveguide.

93. A machine-implemented method for transforming reaction data into modification detection data, wherein the reaction data is representative of a series of events during a sequencing-by-synthesis reaction wherein a nascent strand is synthesized based upon a nucleotide sequence of a template nucleic acid, and the modification detection data is representative of a presence of one or more modifications within a template nucleic acid, the machine-implemented method comprising:

a) creation of a classifier to distinguish between true incorporations and stochastic pulses in an observed signal trace that comprises the true incorporations and the stochastic pulses;
b) slide a fixed-length window across the observed signal trace and use the classifier to count the true incorporations and the stochastic pulses; and
c) based upon the results of step b, identify regions of the observed signal trace in which a density of the stochastic pulses exceeds a density of the true incorporations, wherein the regions are characterized as candidate regions comprising at least one of the one or more modifications.

94. The method of claim 93, wherein steps a-c are performed via a user interface implemented in a machine that comprises instructions stored in machine-readable medium and a processor that executes the instructions.

95. A computer program product comprising a computer usable medium having

a) a computer readable program code embodied therein, said computer readable program code adapted to be executed to implement the method of claim 93; and

b) the machine-readable medium on which the results of steps a-c are stored.

96. A machine-implemented method for transforming reaction data into modification detection data, wherein the reaction data is representative of a series of events during a sequencing-by-synthesis reaction wherein a nascent strand is synthesized based upon a nucleotide sequence of a template nucleic acid, and the modification detection data is representative of a presence of one or more modifications within a template nucleic acid, the machine-implemented method comprising:

a) creation of a classifier to distinguish between true incorporations and stochastic pulses in an observed signal trace that comprises the true incorporations and the stochastic pulses;

b) slide a fixed-length window across the observed signal trace and determine times between the true incorporations; and

c) based upon the results of step b, identify regions of the observed signal trace that have a large difference in time between the true incorporations, wherein the regions are characterized as candidate regions comprising at least one of the one or more modifications.

97. The method of claim 96, wherein steps a-c are performed via a user interface implemented in a machine that comprises instructions stored in machine-readable medium and a processor that executes the instructions.

98. A computer program product comprising a computer usable medium having

a) a computer readable program code embodied therein, said computer readable program code adapted to be executed to implement the method of claim 96; and

b) the machine-readable medium on which the results of steps a-c are stored.

99. A machine-implemented method for transforming reaction data into modification detection data, wherein the reaction data is representative of a series of events during a sequencing-by-synthesis reaction wherein a nascent strand is synthesized based upon a nucleotide sequence of a template nucleic acid, and the modification detection data is representative of a presence of one or more modifications within a template nucleic acid, the

machine-implemented method comprising a segmenting algorithm based on a hidden Markov model architecture.

100. A machine-implemented method for transforming reaction data into modification detection data, wherein the reaction data is representative of a series of events during a sequencing-by-synthesis reaction wherein a nascent strand is synthesized based upon a nucleotide sequence of a template nucleic acid, and the modification detection data is representative of a presence of one or more modifications within a template nucleic acid, the machine-implemented method comprising a segmenting algorithm based on a conditional random field framework.

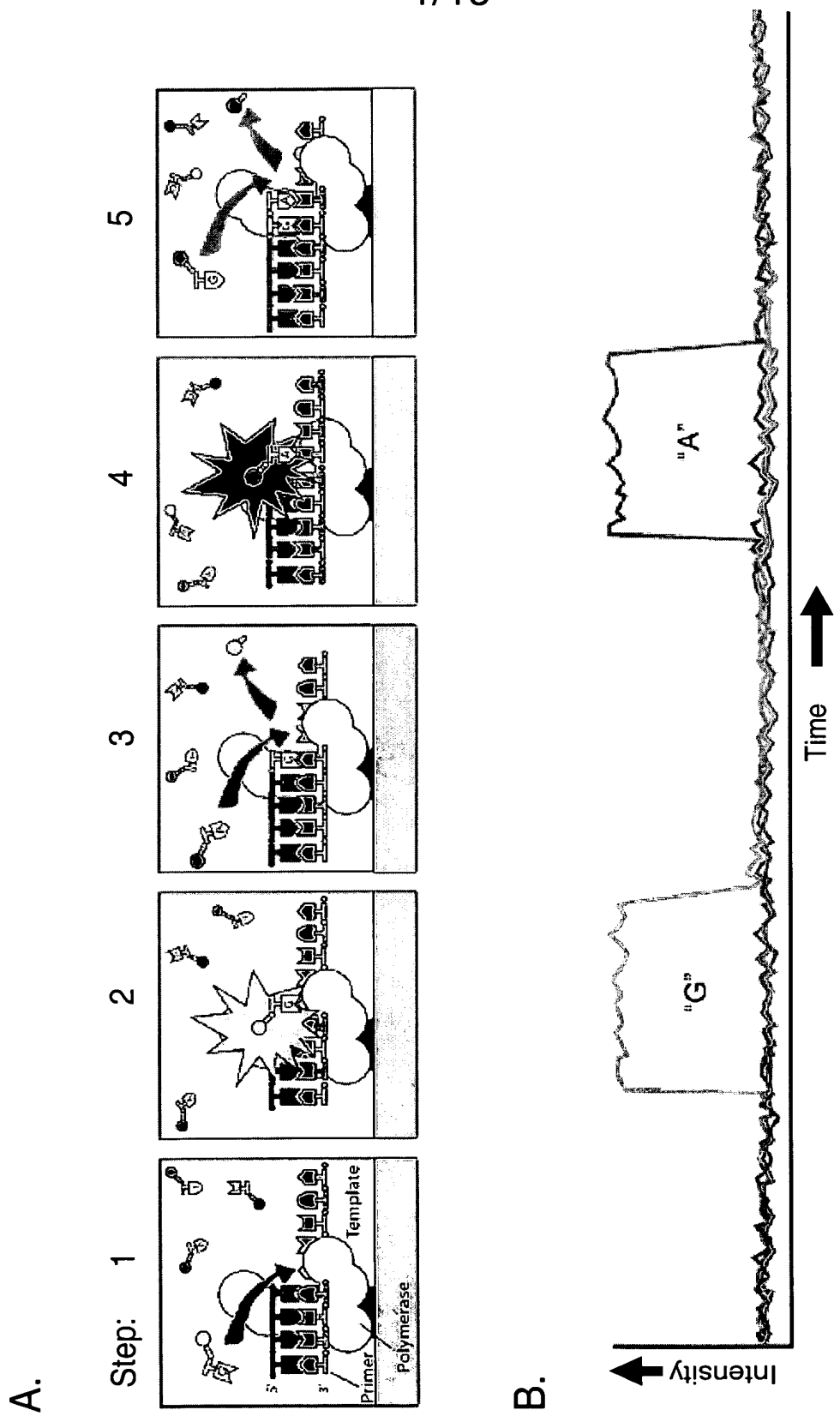

A.
Step: 1
2
3
4
5
Template
Primer
Polymerase
B.
"G"
"A"
Intensity
Time

Figure 1

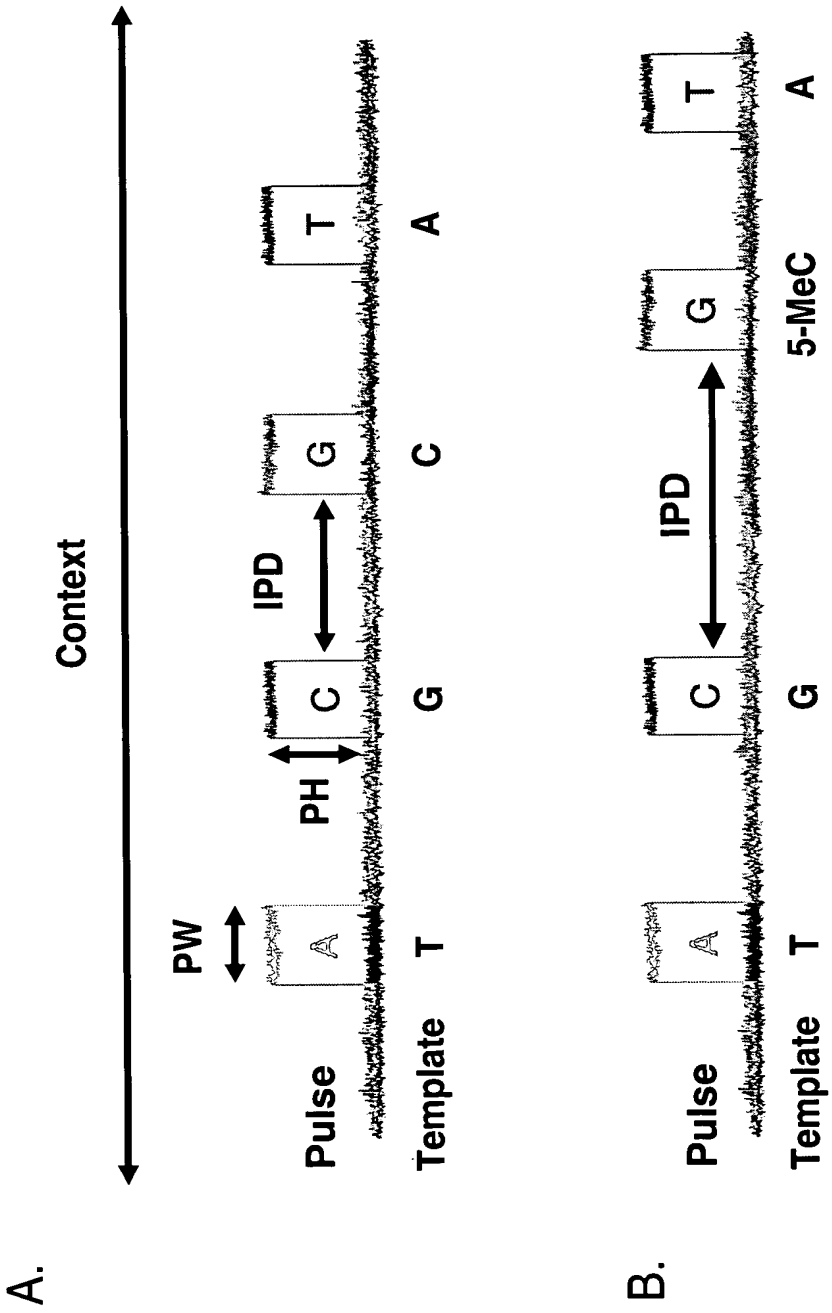

A.
Context
PW
PH
IPD
Pulse
Template
T
A
C
G
T
G
C
A
B.
IPD
Pulse
Template
T
A
C
G
T
G
5-MeC
A

Figure 2

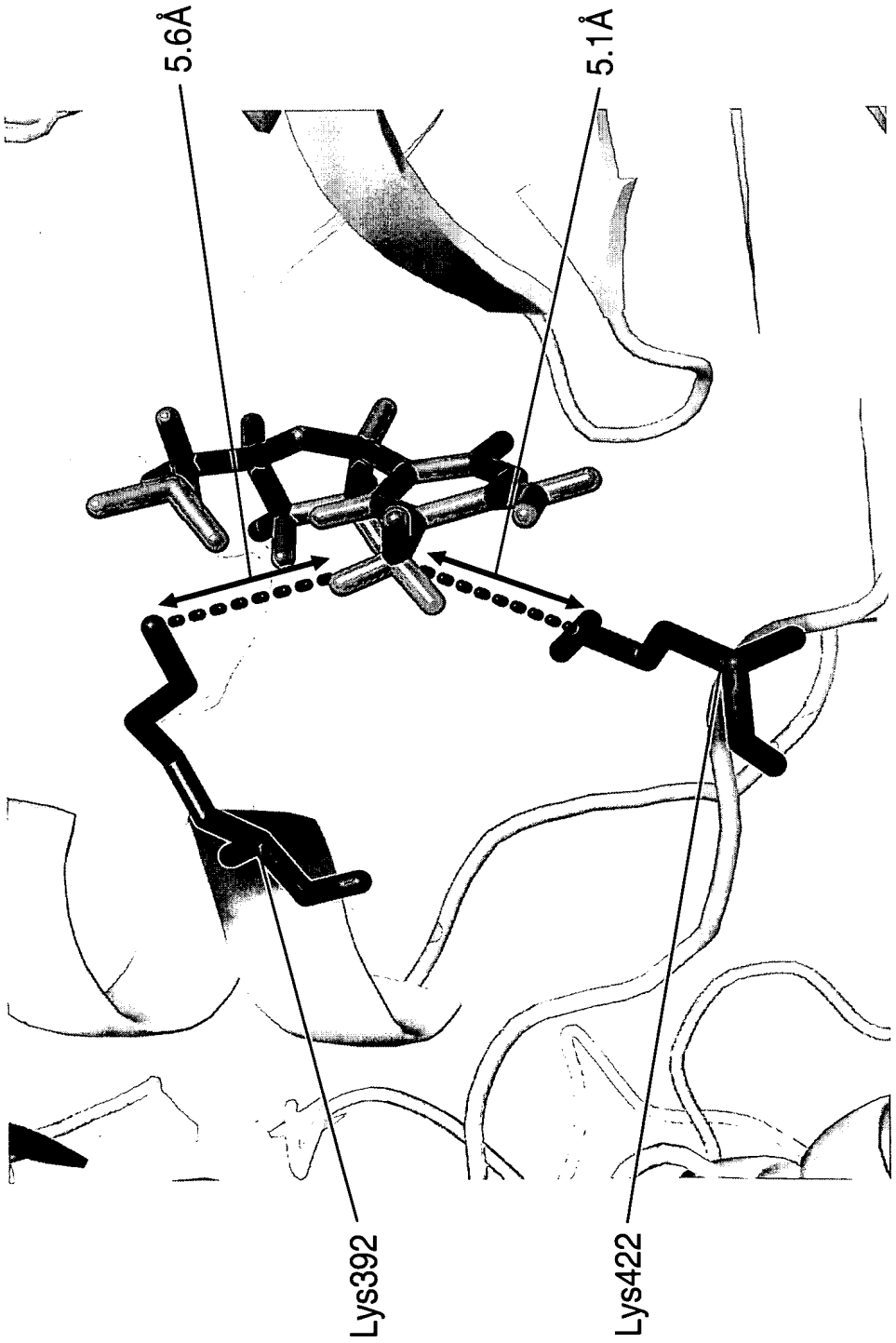


Figure 3

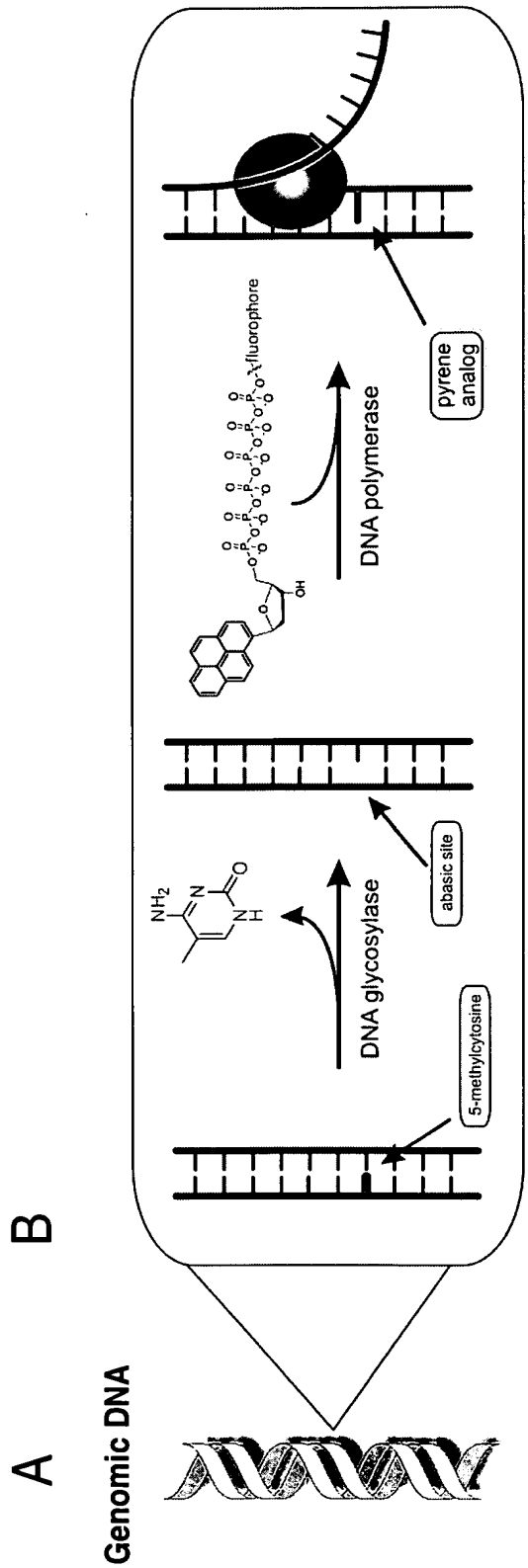

A
Genomic DNA
B
5-methylcytosine
DNA glycosylase
abasic site
DNA polymerase
pyrene analog
γ-fluorophore

Figure 4

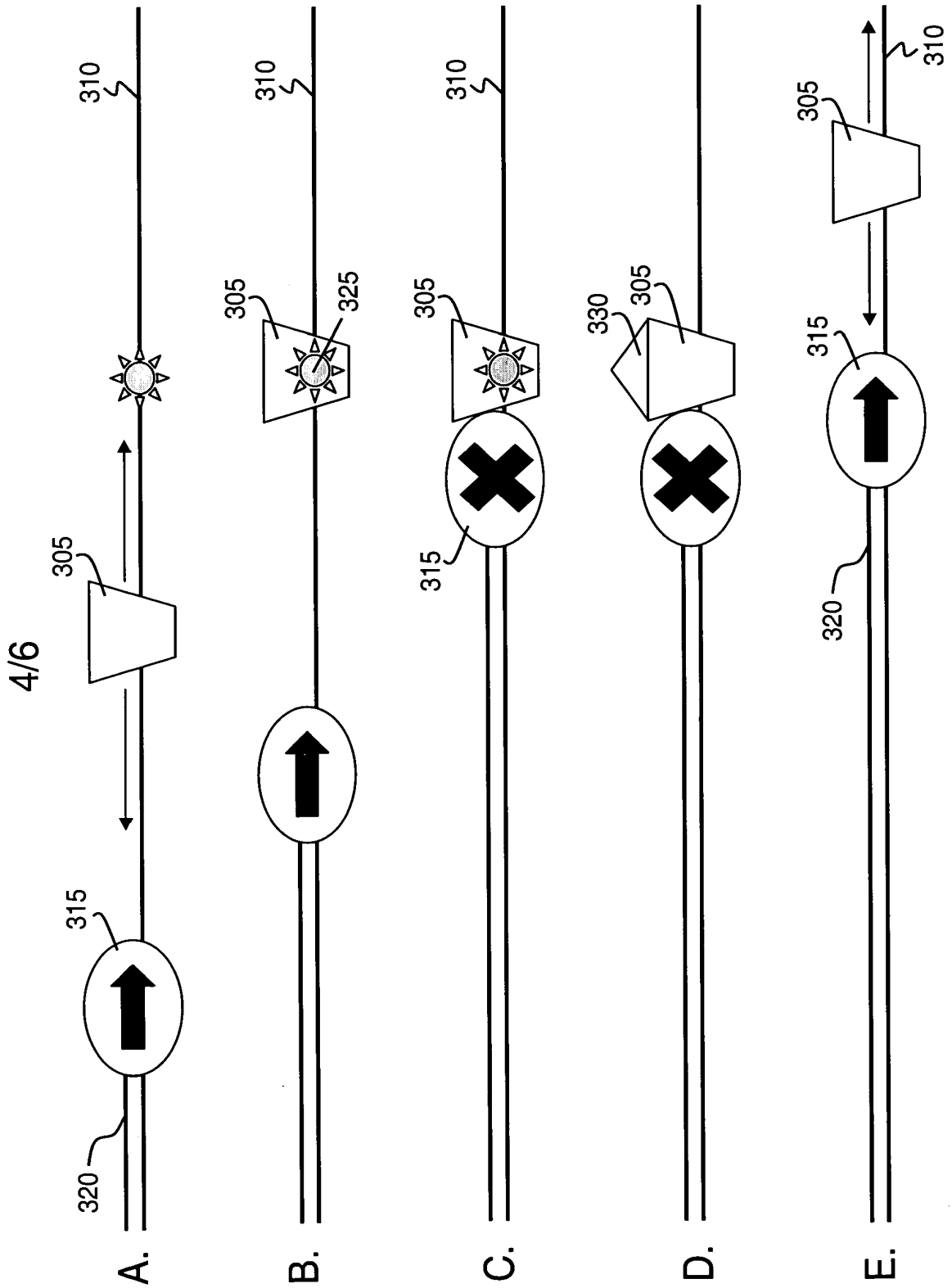


Figure 5

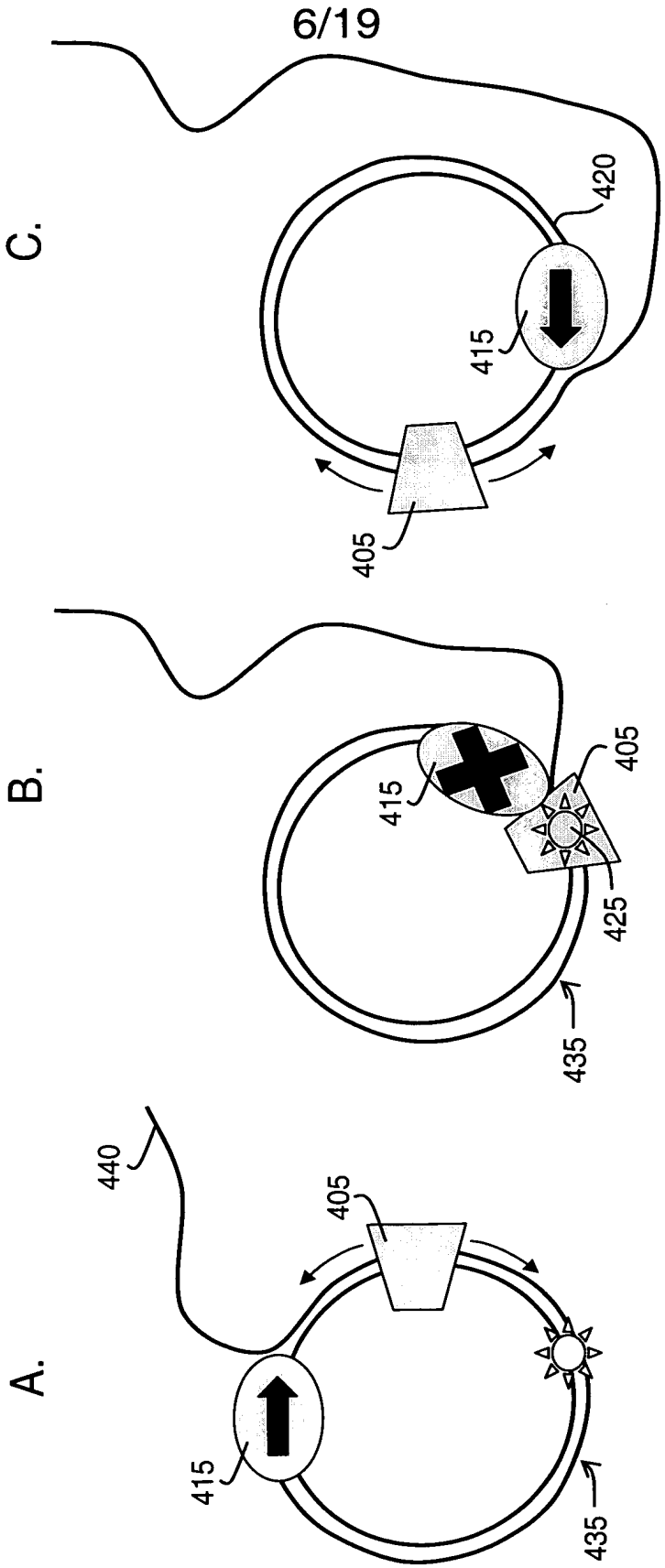


Figure 6

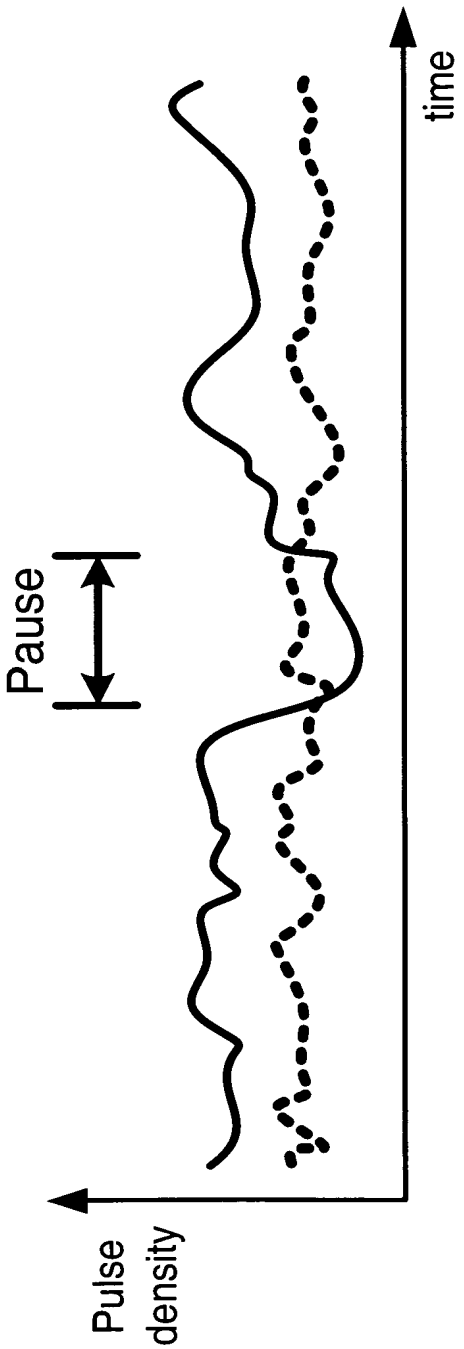


Figure 7

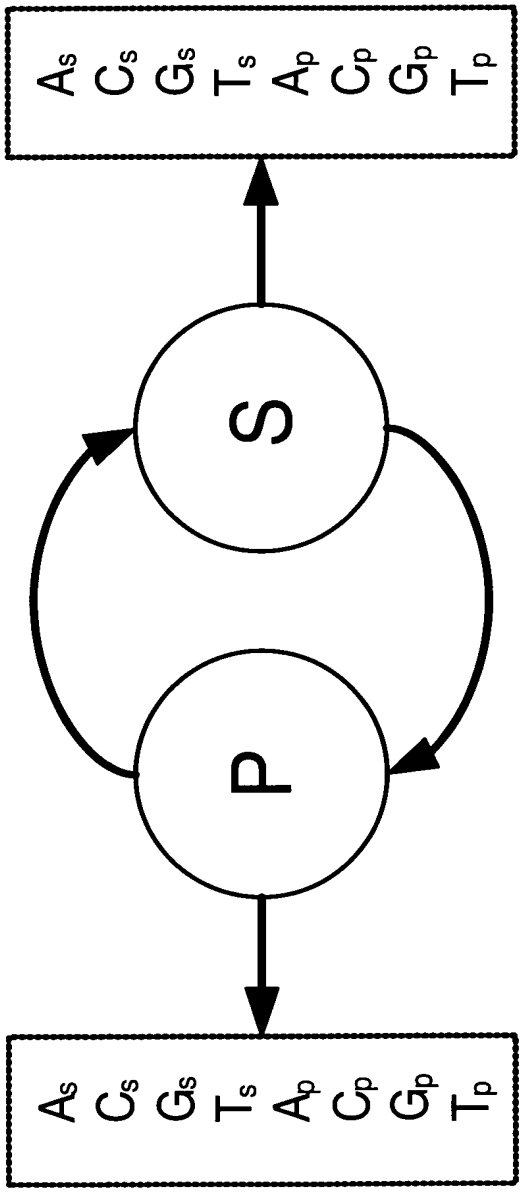


Figure 8

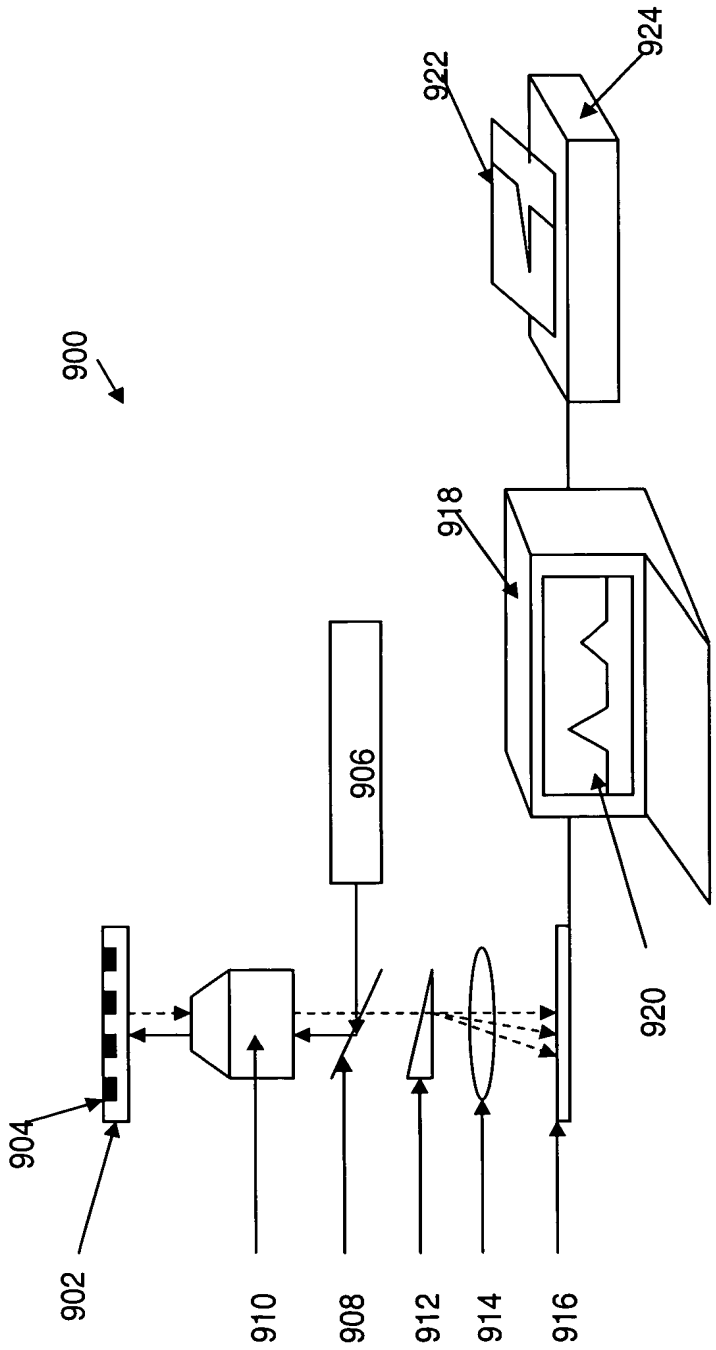


Figure 9

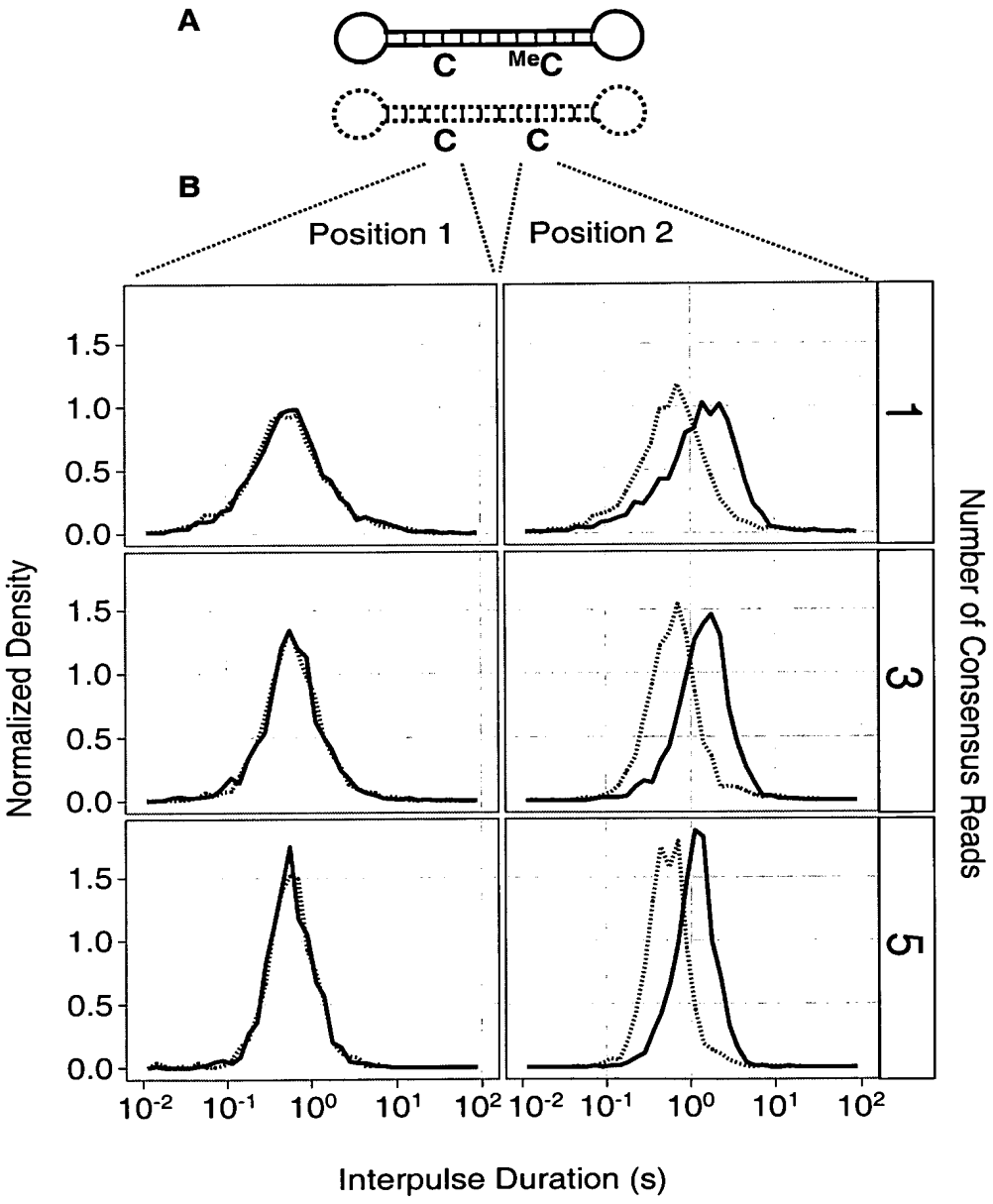

A
C
MeC
C
C
B
Position 1
Position 2
1.5
1.0
0.5
0.0
1
3
5
Normalized Density
Number of Consensus Reads
10^{-2} 10^{-1} 10^{0} 10^{1} 10^{2}
Interpulse Duration (s)

Figure 10

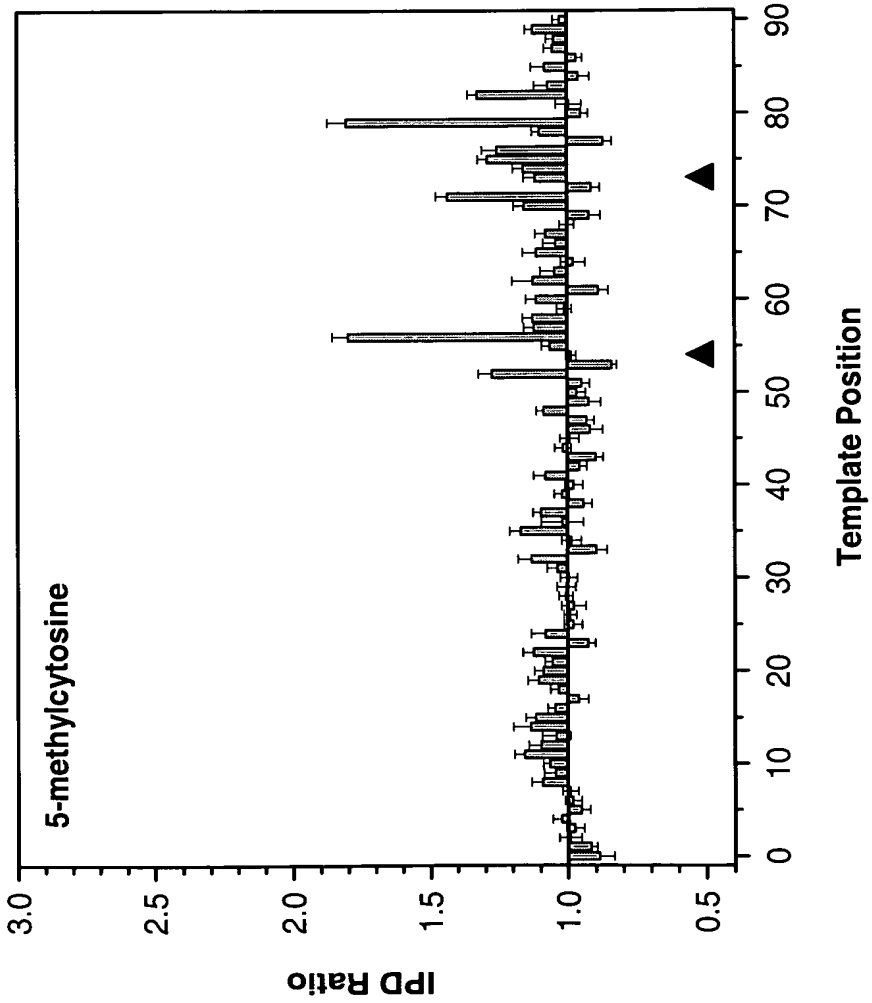

5-methylcytosine
IPD Ratio
3.0
2.5
2.0
1.5
1.0
0.5
0
10
20
30
40
50
60
70
80
90
Template Position

Figure 11

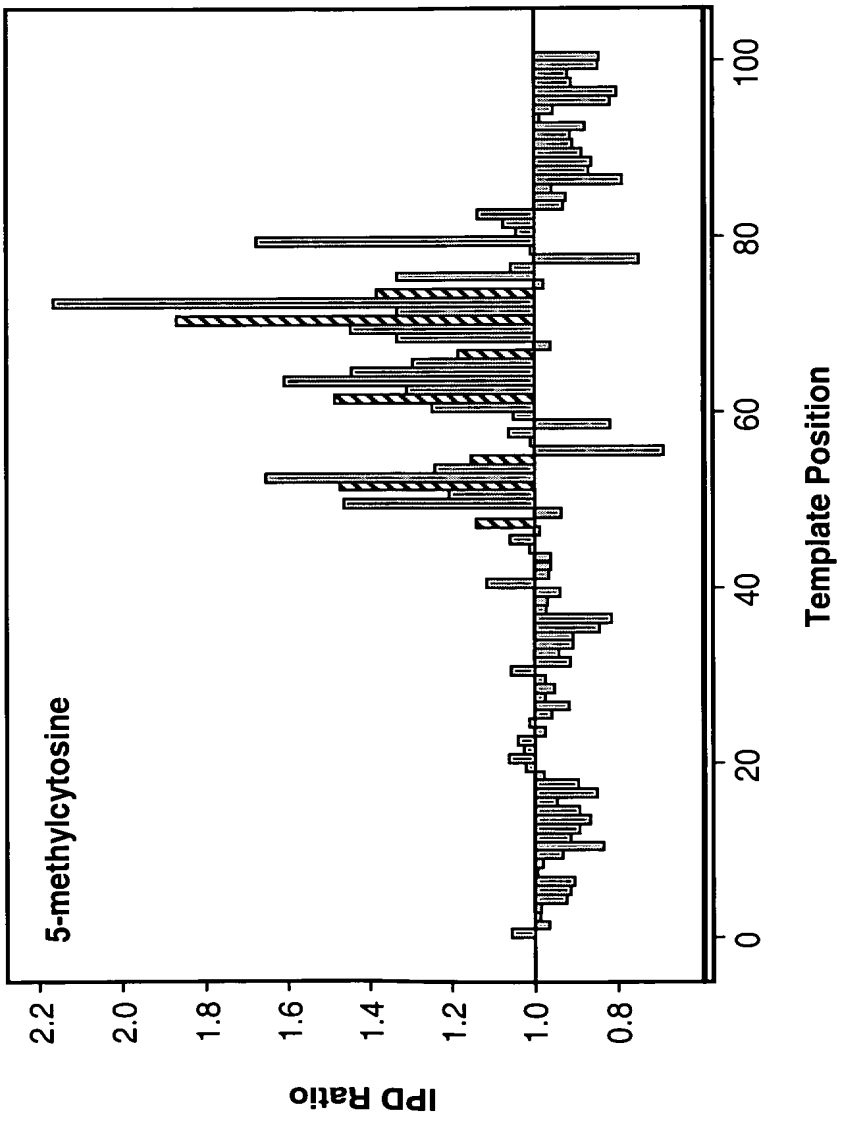

5-methylcytosine
IPD Ratio
2.2
2.0
1.8
1.6
1.4
1.2
1.0
0.8
0
20
40
60
80
100
Template Position

Figure 12

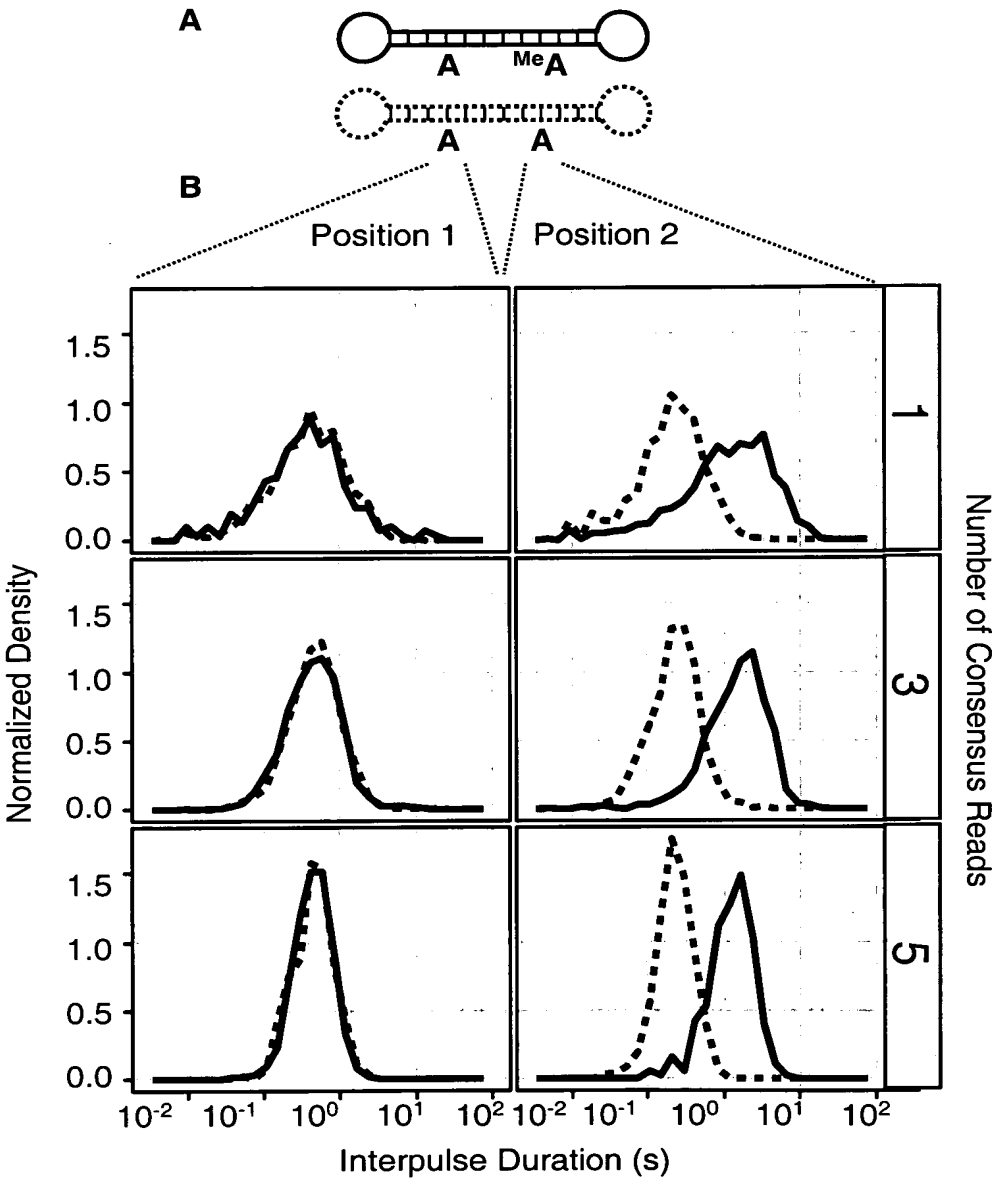

A
A
Me
A
A
A
B
Position 1
Position 2
1.5
1.0
0.5
0.0
1
3
5
Normalized Density
Number of Consensus Reads
10^{-2} 10^{-1} 10^{0} 10^{1} 10^{2}
Interpulse Duration (s)

Figure 13

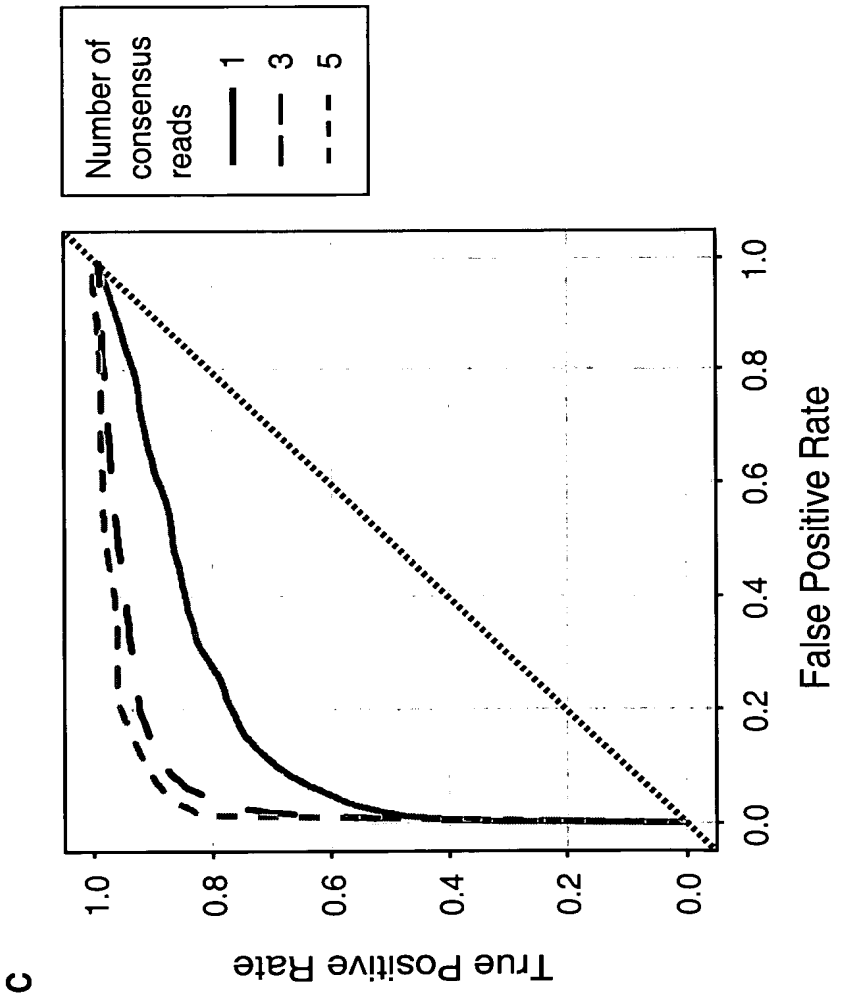


Figure 13, cont.

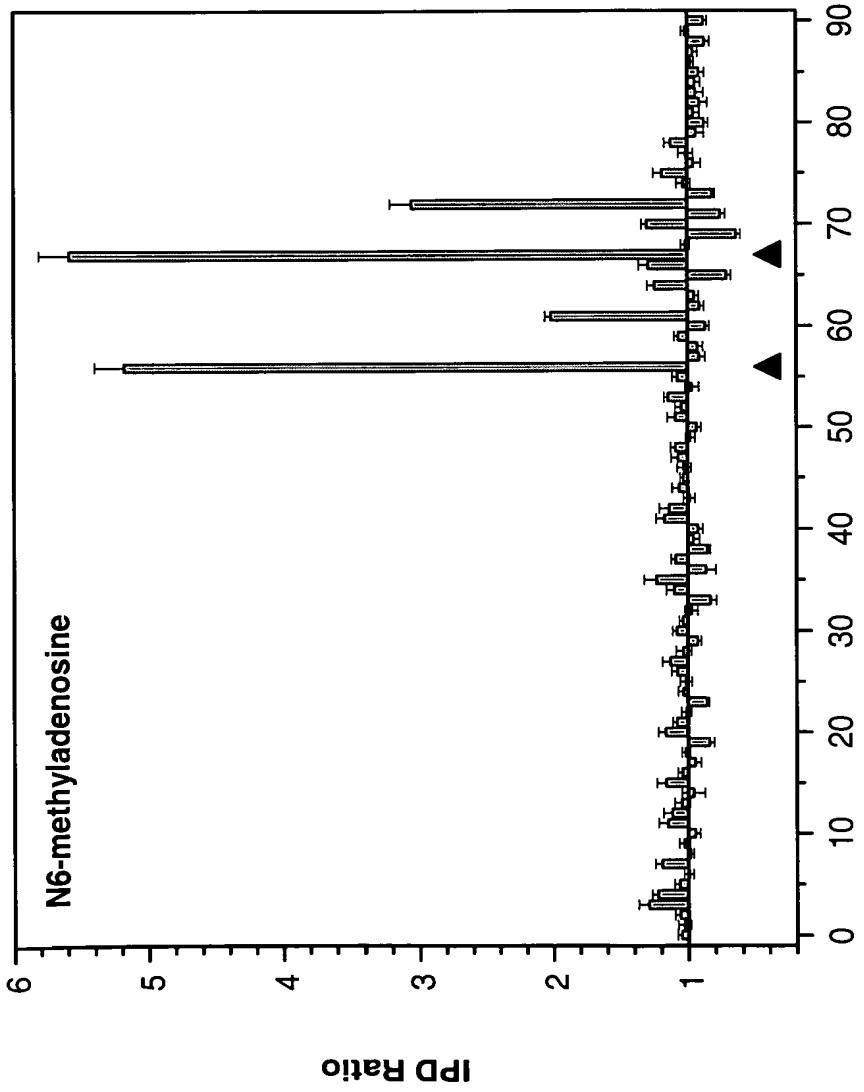

N6-methyladenosine
IPD Ratio
Template Position
0
10
20
30
40
50
60
70
80
90
1
2
3
4
5
6

Figure 14

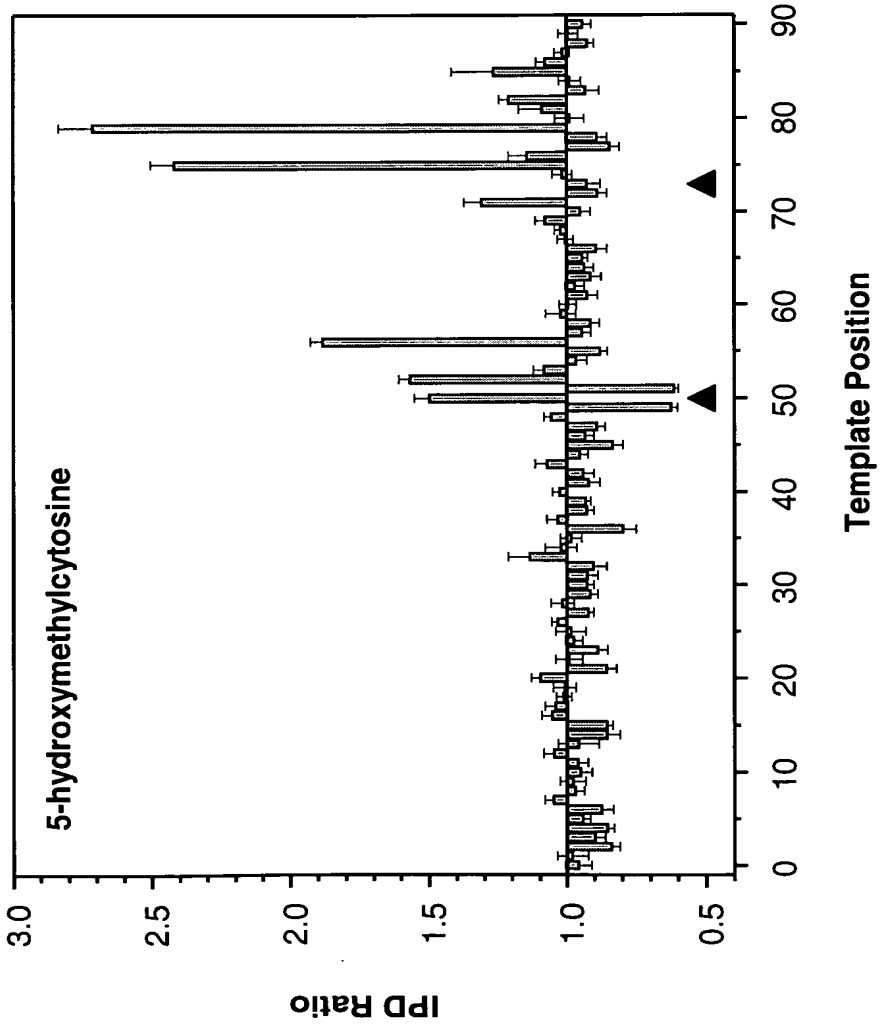

5-hydroxymethylcytosine
IPD Ratio
3.0
2.5
2.0
1.5
1.0
0.5
0
10
20
30
40
50
60
70
80
90
Template Position

Figure 15

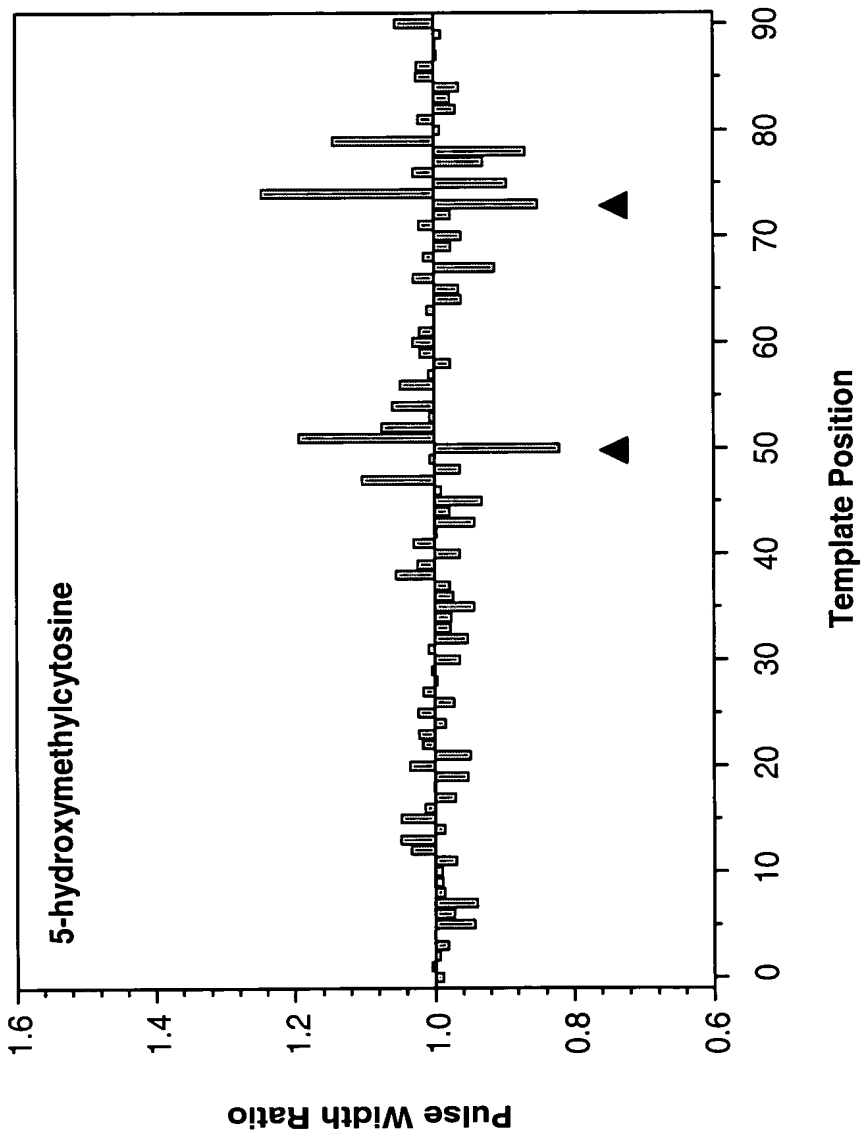

5-hydroxymethylcytosine
Pulse Width Ratio
1.6
1.4
1.2
1.0
0.8
0.6
0
10
20
30
40
50
60
70
80
90
Template Position

Figure 16

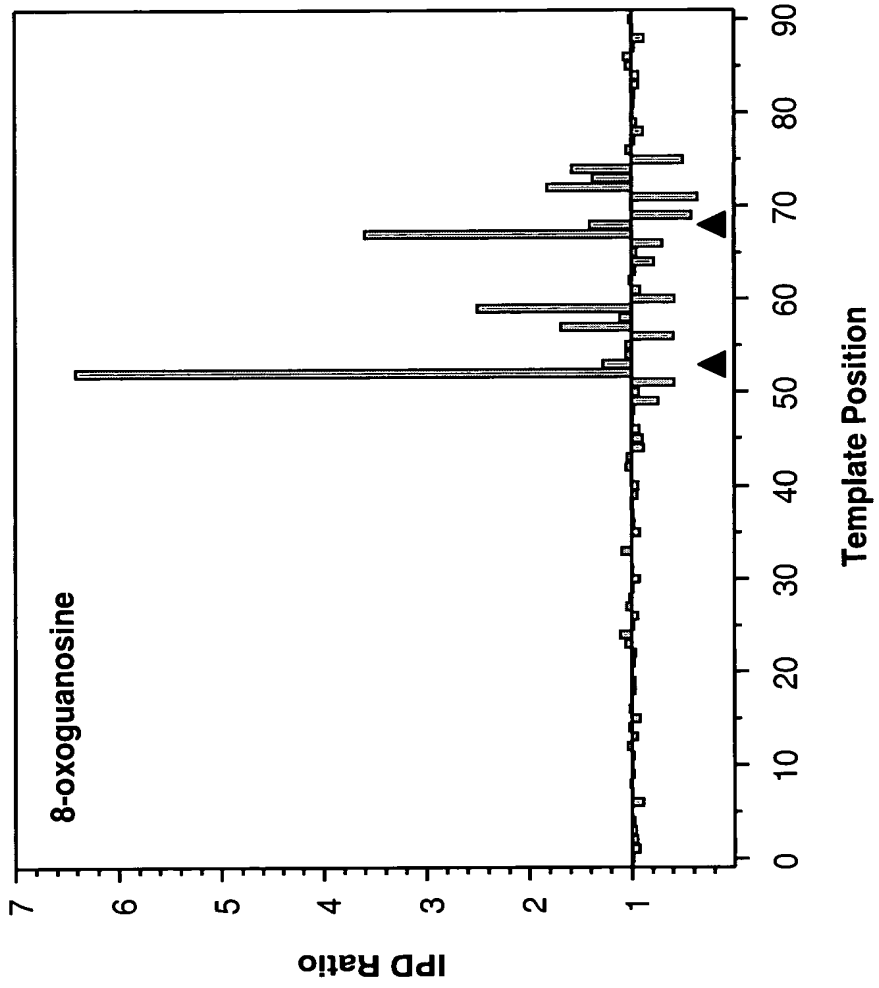

8-oxoguanosine
IPD Ratio
Template Position
1
2
3
4
5
6
7
0
10
20
30
40
50
60
70
80
90

Figure 17

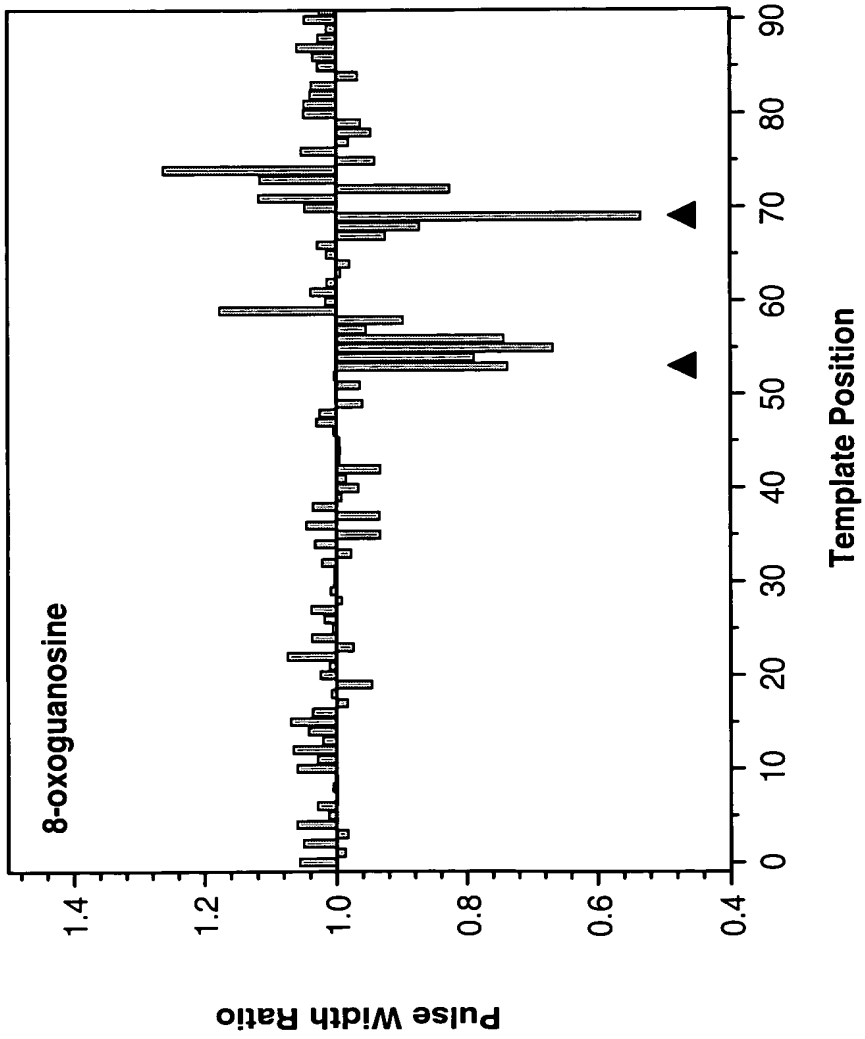

8-oxoguanosine
Pulse Width Ratio
1.4
1.2
1.0
0.8
0.6
0.4
0
10
20
30
40
50
60
70
80
90
Template Position

Figure 18

Espacenet

Bibliographic data: WO2013170429 (A1) — 2013-11-21

METHOD, SYSTEM AND COMPUTER READABLE MEDIUM FOR DETERMINING BASE INFORMATION IN PREDETERMINED AREA OF FETUS GENOME

Inventor(s): CHEN SHENGPEI [CN]; GE HUIJUAN [CN]; LI XUCHAO [CN]; YI SHANG [CN]; WANG JIAN [CN]; WANG JUN [CN]; YANG HUANMING [CN]; ZHANG XIUQING [CN] ± (CHEN, SHENGPEI, ; GE, HUIJUAN, ; LI, XUCHAO, ; YI, SHANG, ; WANG, JIAN, ; WANG, JUN, ; YANG, HUANMING, ; ZHANG, XIUQING)

Applicant(s): BGI HEALTH SERVICE CO LTD [CN]; CHEN SHENGPEI [CN]; GE HUIJUAN [CN]; LI XUCHAO [CN]; YI SHANG [CN]; WANG JIAN [CN]; WANG JUN [CN]; YANG HUANMING [CN]; ZHANG XIUQING [CN] ± (BGI HEALTH SERVICE CO., LTD, ; CHEN, SHENGPEI, ; GE, HUIJUAN, ; LI, XUCHAO, ; YI, SHANG, ; WANG, JIAN, ; WANG, JUN, ; YANG, HUANMING, ; ZHANG, XIUQING)

Classification: - international: *C12Q1/68*
- cooperative: **C12Q1/6809 (CN)**; **C12Q1/6858 (EP, KR, US)**; **C12Q1/6869 (EP, KR, US)**; **C12Q1/6883 (US)**; **G06F19/18 (KR)**; **G06F19/22 (KR)**; **G06F19/24 (KR)**; **G06F19/3431 (KR)**; **G06F19/345 (KR)**; **G16B20/00 (EP, US)**; **G16B30/00 (EP, US)**; **G16H50/20 (EP, US)**; **G16H50/30 (EP, US)**; C12Q2600/118 (US); C12Q2600/156 (US) more

Application number: WO2012CN75478 20120514 Global Dossier

Priority number(s): WO2012CN75478 20120514

Also published as: AU2012380221 (A1) AU2012380221 (B2) CN104053789 (A) CN104053789 (B) EP2851431 (A1) EP2851431 (A4) EP2851431 (B1) ES2656023 (T3) HK1196401 (A1) JP2015525062 (A) JP6045686 (B2) KR101770460 (B1) KR20140146193 (A) PL2851431 (T3) RU2014150655 (A) RU2597981 (C2) SG11201407515R (A) US2015094210 (A1) US2018320235 (A1) less

Abstract of WO2013170429 (A1)

Provided are a method, system and computer readable medium for determining the base information in a predetermined area of a fetus genome, the method comprising the following steps: constructing a sequence library for the DNA samples of the fetus genome; sequencing the sequence library to obtain the sequencing result of the fetus, the sequencing result of the fetus comprised of a plurality of sequencing data; and based on the sequencing result of the fetus, determining the base information in the predetermined area according to the hidden Markov model in conjunction with the genetic information of an individual related hereditarily to the fetus.

(12) 按照专利合作条约所公布的国际申请

(19) 世界知识产权组织
国 际 局

(43) 国际公布日
2013 年 11 月 21 日 (21.11.2013)

WIPO | PCT

(10) 国际公布号
WO 2013/170429 A1

(51) 国际专利分类号:
C12Q 1/68 (2006.01)

(21) 国际申请号: PCT/CN2012/075478

(22) 国际申请日: 2012 年 5 月 14 日 (14.05.2012)

(25) 申请语言: 中文

(26) 公布语言: 中文

(71) 申请人 (对除美国外的所有指定国): **深圳华大基因健康科技有限公司 (BGI HEALTH SERVICE CO., LTD.)** [CN/CN]; 中国广东省深圳市盐田区北山道 146 号北山工业区 11 栋 2、3 楼, Guangdong 518083 (CN)。

(72) 发明人; 及

(75) 发明人/申请人 (仅对美国): **陈盛培 (CHEN, Shengpei)** [CN/CN]; 中国广东省深圳市盐田区北山工业区综合楼, Guangdong 518083 (CN)。 **葛会娟 (GE, Huijuan)** [CN/CN]; 中国广东省深圳市盐田区北山工业区综合楼, Guangdong 518083 (CN)。 **李旭超 (LI, Xuchao)** [CN/CN]; 中国广东省深圳市盐田区北山工业区综合楼, Guangdong 518083 (CN)。 **易赏 (YI, Shang)** [CN/CN]; 中国广东省深圳市盐田区北山工业区综合楼, Guangdong 518083 (CN)。 **汪建 (WANG, Jian)** [CN/CN]; 中国广东省深圳市盐田区北山工业区综合楼, Guangdong 518083 (CN)。 **王俊 (WANG, Jun)** [CN/CN]; 中国广东省深圳市盐田区北山工业区综合楼, Guangdong 518083 (CN)。 **杨焕明 (YANG, Huanming)** [CN/CN]; 中国广东省深圳市盐田区北山工业区综合楼, Guangdong 518083 (CN)。

张秀清 (ZHANG, Xiuqing) [CN/CN]; 中国广东省深圳市盐田区北山工业区综合楼, Guangdong 518083 (CN)。

(74) 代理人: **北京清亦华知识产权代理事务所（普通合伙） (TSINGYIHUA INTELLECTUAL PROPERTY LLC)**; 中国北京市海淀区清华园清华大学照澜院商业楼 301 室, Beijing 100084 (CN)。

(81) 指定国 (除另有指明，要求每一种可提供的国家保护): 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, QA, RO, RS, RU, RW, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW。

(84) 指定国 (除另有指明，要求每一种可提供的地区保护): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, SZ, TZ, UG, ZM, ZW), 欧亚 (AM, AZ, BY, KG, KZ, RU, TJ, TM), 欧洲 (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG)。

本国际公布:

— 包括国际检索报告(条约第 21 条(3))。

(54) **Title**: METHOD, SYSTEM AND COMPUTER READABLE MEDIUM FOR DETERMINING BASE INFORMATION IN PREDETERMINED AREA OF FETUS GENOME

(54) **发明名称** : 确定胎儿基因组中预定区域碱基信息的方法、系统和计算机可读介质

(57) **Abstract**: Provided are a method, system and computer readable medium for determining the base information in a predetermined area of a fetus genome, the method comprising the following steps: constructing a sequence library for the DNA samples of the fetus genome; sequencing the sequence library to obtain the sequencing result of the fetus, the sequencing result of the fetus comprised of a plurality of sequencing data; and based on the sequencing result of the fetus, determining the base information in the predetermined area according to the hidden Markov model in conjunction with the genetic information of an individual related hereditarily to the fetus.

(57) **摘要**: 提供了确定胎儿基因组中预定区域碱基信息的方法、系统和计算机可读介质。其中，确定胎儿基因组中预定区域碱基信息的方法，包括下列步骤：针对胎儿基因组 DNA 样本，构建测序文库；对测序文库进行测序，以便获得胎儿的测序结果，该胎儿的测序结果由多个测序数据构成；基于胎儿的测序结果，结合胎儿遗传相关个体的遗传信息，根据隐马尔可夫模型，确定预定区域的碱基信息。

WO 2013/170429 A1

确定胎儿基因组中预定区域碱基信息的方法、系统和计算机可读介质

优先权信息

无

技术领域

本发明涉及确定胎儿基因组中预定区域碱基信息的方法、系统和计算机可读介质。

背景技术

遗传性疾病是由于遗传物质发生改变而造成的疾病，具有先天性、家族性、终身性和遗传性的特点。遗传性疾病可分为 3 个大类：单基因遗传病、多基因遗传病及染色体异常。其中单基因病多由于单个致病基因的显性或隐性遗传所致基因功能异常；而多基因遗传病则是由多个基因变化影响所致的疾病，会在一定程度上受到外界环境因素的影响；染色体异常包括数目异常和结构异常，最为多见的是由于第 21 号染色体三体所致的唐氏综合症，患儿表现为先天愚型和肢体形状异常等其他先天性特征。由于目前对遗传性疾病尚无有效的治疗方式，只能针对性地进行支持治疗或者药物缓解，费用昂贵，给社会和家庭带来沉重经济和精神负担。因此，在孩子出生前就对孩子的患病状态进行检测，做好预防工作，以达到优生优育的目的，是十分必要的。

然而，目前的相关检测手段仍有待改进。

发明内容

本发明旨在至少解决现有技术中存在的技术问题之一。

在本发明的一个方面，本发明提出了一种确定胎儿基因组中预定区域碱基信息的方法。根据本发明的实施例，该方法包括下列步骤：针对胎儿基因组 DNA 样本，构建测序文库；对所述测序文库进行测序，以便获得胎儿的测序结果，所述胎儿的测序结果由多个测序数据构成；基于所述胎儿的测序结果，结合胎儿遗传相关个体的遗传信息，根据隐马尔可夫模型，确定所述预定区域的碱基信息。子代的基因组形成过程，相当于亲代基因组的一次随机重组（即连锁互换单倍体型重组，以及配子的随机组合）。对于孕期血浆，假若我们将胎儿的单倍型（父母单倍型的重组型）作为隐含状态（hidden states），可以将血浆的测序数据当做观察序列（observations），借助先验数据推算出状态转移概率（transition probabilities）、观察序列概率分布（observation symbol probabilities）和初始状态概率分布（initial state distribution），我们则可以通过诸如惠特比算法（Viterbi algorithm）根据隐马尔可夫模型推断出最可能的胎儿单倍型组合，从而获得更多胎儿的信息。因而，根据本发明的实施例，借助隐马尔可夫模型，例如可以通过利用惠特比算法（Viterbi algorithm），参考胎儿遗传相关个体的遗传信息，可以确定胎儿基因组中特定区域的核酸序列，由此，可以有效地对胎儿基因组的遗传信息进行产前检测。

在本发明的又一方面，本发明提出了一种用于确定胎儿基因组中预定区域碱基信息的系统。根据本发明的实施例，该系统包括：文库构建装置，所述文库构建装置适于针对胎儿基因组DNA样本，构建测序文库；测序装置，所述测序装置与所述文库构建装置相连，并且适于对所述测序文库进行测序，以便以便获得胎儿的测序结果，所述胎儿的测序结果由多个测序数据构成；分析装置，基于所述胎儿的测序结果，结合胎儿遗传相关个体的遗传信息，根据隐马尔可夫模型，确定所述预定区域的碱基信息。利用该系统，能够有效地实施前面所述的确定胎儿基因组中预定区域碱基信息的方法，可以借助隐马尔可夫模型，例如可以通过利用惠特比算法（Viterbi algorithm），参考胎儿遗传相关个体的遗传信息，可以确定胎儿基因组中特定区域的核酸序列，由此，可以有效地对胎儿基因组的遗传信息进行产前检测，从而可以有效地对胎儿基因组的遗传信息进行产前确定。

在本发明的另一方面，本发明还提出了一种计算机可读介质。根据本发明的实施例，该计算机可读介质上存储有指令，所述指令适于被处理器执行以便基于胎儿的测序结果，结合胎儿遗传相关个体的遗传信息，根据隐马尔可夫模型，确定预定区域的碱基信息。利用本发明的计算机可读介质，能够有效地被处理器执行其存储的指令，以便借助隐马尔可夫模型，例如可以通过利用惠特比算法（Viterbi algorithm），基于胎儿的测序结果，参考胎儿遗传相关个体的遗传信息，可以确定胎儿基因组中特定区域的核酸序列，由此，可以有效地对胎儿基因组的遗传信息进行产前检测。

本发明的附加方面和优点将在下面的描述中部分给出，部分将从下面的描述中变得明显，或通过本发明的实践了解到。

附图说明

本发明的上述和/或附加的方面和优点从结合下面附图对实施例的描述中将变得明显和容易理解，其中：

图1为根据本发明一个实施例的利用隐马尔可夫模型进行分析的流程示意图；以及

图2为根据本发明的一个实施例的用于确定胎儿基因组中预定区域核酸序列的系统的结构示意图。

发明详细描述

下面详细描述本发明的实施例，所述实施例的示例在附图中示出，其中自始至终相同或类似的标号表示相同或类似的元件或具有相同或类似功能的元件。下面通过参考附图描述的实施例是示例性的，仅用于解释本发明，而不能理解为对本发明的限制。

需要说明的是，术语“第一”、“第二”仅用于描述目的，而不能理解为指示或暗示相对重要性或者隐含指明所指示的技术特征的数量。由此，限定有“第一”、“第二”的特征可以明示或者隐含地包括一个或者更多个该特征。进一步地，在本发明的描述中，除非另有说明，“多个”的含义是两个或两个以上。

确定胎儿基因组中预定区域碱基信息的方法

在本发明的第一方面，本发明提出了一种确定胎儿基因组中预定区域碱基信息的方法。根据本发明的实施例，该方法包括下列步骤：

首先，针对胎儿基因组 DNA 样本，构建测序文库。根据本发明的实施例，胎儿基因组 DNA 样本的来源并不受特别限制。根据本发明的一些实施例，可以采用任何含有胎儿核酸的孕妇样本。例如，根据本发明的实施例，可以采用的孕妇样本为孕妇乳汁、尿液和外周血。其中，优选孕妇外周血。采用孕妇外周血作为胎儿基因组 DNA 样本的来源，可以有效地实现无创取样方式获得胎儿基因组 DNA，从而可以在不影响胎儿正常发育的前提下，对胎儿的基因组进行有效监测。关于针对核酸样本，构建测序文库的方法和流程，本领域技术人员可以根据不同的测序技术进行适当选择，关于流程的细节，可以参见测序仪器的厂商例如 Illumina 公司所提供的规程，例如参见 Illumina 公司 Multiplexing Sample Preparation Guide（Part#1005361; Feb 2010）或 Paired-End SamplePrep Guide（Part#1005063; Feb 2010），通过参照将其并入本文。根据本发明的实施例，从生物样本提取核酸样本的方法和设备，也不受特别限制，可以采用商品化的核酸提取试剂盒进行。

在构建测序文库后，将测序文库应用于测序仪器，对测序文库进行测序，并获得相应的测序结果，该测序结果是由多个测序数据构成的。根据本发明的实施例，可以用于进行测序的方法和设备并不受特别限制，包括但不限于双脱氧链终止法；优选高通量的测序方法，由此，能够利用这些测序装置的高通量、深度测序的特点，进一步提高测序效率。从而，能够提高后续对测序数据进行分析，尤其是统计检验分析时的精确性和准确度。所述高通量的测序方法包括但不限于第二代测序技术或者是单分子测序技术。所述第二代测序平台（Metzker ML. Sequencing technologies-the next generation. Nat Rev Genet. 2010 Jan;11(1):31-46）包括但不限于 Illumina-Solexa（GA™,HiSeq2000™ 等）、ABI-Solid 和 Roche-454（焦磷酸测序）测序平台；单分子测序平台（技术）包括但不限于 Helicos 公司的真实单分子测序技术（True Single Molecule DNA sequencing）， Pacific Biosciences 公司单分子实时测序（single molecule real-time (SMRT™)），以及 Oxford Nanopore Technologies 公司的纳米孔测序技术等（Rusk, Nicole (2009-04-01). Cheap Third-Generation Sequencing. Nature Methods 6 (4): 244-245）。随着测序技术的不断进化，本领域技术人员能够理解的是还可以采用其他的测序方法和装置进行全基因组测序。根据本发明的具体示例，可以利用选自 Illumina-Solexa、ABI-SOLiD、Roche-454 和单分子测序装置的至少一种对所述全基因组测序文库进行测序。

任选地，在得到测序结果之后，可以将所述测序结果与参照序列进行比对，以便确定与所述预定区域对应的测序数据。在本文中所使用的术语“预定区域”应作广义理解，是指任何包含可能发生预定事件位点的核酸分子的区域。对于 SNP 分析而言，可以是指包含 SNP 位点的区域。对于分析染色体非整倍性，则预定区域指的是所要分析的染色体的全长或者部分，即选择所有来自该染色体的测序数据。从测序结果中选择来自相应区域的测序数据的方法可以不受特别限制。根据本发明的实施例，可以通过将所得到的所有测序数据与已知的核酸参照序列进行比对，从而得到来自于预定区域的测序数据。另外，根据本发

明的实施例，预定区域也可以是基因组上不连续的多个分散点。根据本发明的实施例，可以使用的参照序列的类型并不受特别限制，可以为任何含有感兴趣区域的已知序列。根据本发明的实施例，可以采用已知的人类参考基因组作为参照序列。例如，根据本发明的实施例，采用的人类参考基因组为 NCBI 36.3，HG18。另外，根据本发明的实施例，进行比对的方法并不受特别限制。根据本发明的具体实施例，可以采用 SOAP 进行比对。

接下来，基于与预定区域对应的测序数据，确定预定区域中的部分核酸序列；以及基于所确定的预定区域中的部分核酸序列，按照惠特比算法，确定预定区域的其他核酸序列，以便获得预定区域的核酸序列。根据本发明的实施例，可以通过基于所述胎儿的测序结果，结合胎儿遗传相关个体的遗传信息，根据隐马尔可夫模型，确定所述预定区域的碱基信息。根据本发明的实施例，可以利用惠特比算法（Viterbi algorithm），借助隐马尔可夫模型，确定胎儿基因组中特定区域的碱基信息。由此，可以有效地对胎儿基因组的遗传信息进行产前检测。

下面参考图 1，对利用惠特比算法借助隐马尔可夫模型进行分析的原理进行详细描述：

在本文中所使用的术语“胎儿遗传相关个体”指的是在遗传意义上，与胎儿之间具有亲缘关系的个体，例如根据本发明的实施例，可以采用的“胎儿遗传相关个体”为胎儿的亲代例如父母。由此，子代的基因组形成过程，相当于亲代基因组的一次随机重组（即连锁互换单倍体型重组，以及配子的随机组合）。对于孕期血浆，假若我们将胎儿的单倍型（父母单倍型的重组型）作为隐含状态（hidden states），可以将血浆的测序数据当做观察序列（observations），借助先验数据推算出状态转移概率（transition probabilities）、观察序列概率分布（observation symbol probabilities）和初始状态概率分布（initial state distribution），我们则可以通过惠特比算法（Viterbi algorithm）推断出最可能的胎儿单倍型组合，从而获得更多胎儿的信息。

详细分析步骤如下：

记号：

I.需要检测的位点数为 N。

II. 父母的单倍型分别记为 $FH=\{fh_0,fh_1\}$ 和 $MH=\{mh_0,mh_1\}$，

其中，

$mh_k=\{m_{1,k},...,m_{i,k},...,m_{N,k}\}$，$fh_k=\{f_{1,k},...,f_{i,k},...,f_{N,k}\}$，

$\forall fh_{i,k},mh_{i,k}\in\{A,C,G,T\}$，

$k\in\{0,1\}$，$i=1,2,3,...,N$。

III. 将未知的胎儿单倍型记为 $H=\{h_0,h_1\}$，特别地，h_0 和 h_1 分别遗传自母亲和父亲。

$h_0=\{m_{1,x_1},...,m_{i,x_i},...,m_{N,x_N}\}$，$h_1=\{f_{1,y_1},...,f_{i,y_i},...,f_{N,y_N}\}$，

其中， $x_i \in \{0,1\}$， $y_i \in \{0,1\}$，

下标 x_i 和 y_i 组成的序列对， $q_i = \{x_i, y_i\}$ 组成了我们需要解码的隐藏状态，

而所有可能出现的隐藏状态组成了集合 Q。

IV. 测序数据记为， $S = \{s_1, ..., s_i, ..., s_N\}$，

其中， $s_i = \{n_{i,A}, n_{i,C}, n_{i,G}, n_{i,G}\}$，代表此位点的测序信息，包含了 ACGT 四种碱基的数量。

V. 平均胎儿浓度和平均测序错误率分别记为 ε 和 e。

第一步，构建初始状态概率分布向量，以及单倍体重组转移矩阵：

I. 初始状态概率分布记为 $\pi = \{\pi_j\}$ （ $j \in Q$ ）。

根据本发明的实施例，在没有参考数据的情况下，可以设 $\pi_j = \Pr(q_1 = j) \triangleq \frac{1}{4}$，即每种隐藏状态在第一个位点出现的可能性相等。

II. 根据本发明的实施例， 记单倍体重组概率为 $p_r = re/N$，其中 re 代表人类配子基因组重组平均次数，为先验数据在 25 到 30 之间。

III. 根据本发明的实施例，记单倍体重组转移矩阵记为 $A = \{a_{jk}\}$ （ $j,k \in Q$ ），其中 a_{jk} 为隐藏状态转移的概率，即

$$a_{jk} = \Pr(q_i = k | q_{i-1} = j) = \begin{cases} (1-p_r)^2 & x_i = x_{i-1}, y_i = y_{i-1} \\ (1-p_r) \cdot p_r & x_i = x_{i-1}, y_i \neq y_{i-1} \quad or \quad x_i \neq x_{i-1}, y_i = y_{i-1} \\ p_r^2 & x_i \neq x_{i-1}, y_i \neq y_{i-1} \end{cases},$$

胎儿单倍型 $h_0 = \{m_{1,x_1}, ..., m_{i,x_i}, ..., m_{N,x_N}\}$， $h_1 = \{f_{1,y_1}, ..., f_{i,y_i}, ..., f_{N,y_N}\}$ 的下标 x_i 和 y_i 组成的序列对， $q_i = \{x_i, y_i\}$ 组成了我们需要解码的隐藏状态。举个例子， $x_i = 0$ 代表，“母源性染色体上，对应基因座上等位基因型为 $m_{i,0}$ ”。

第二步，构建观察序列概率矩阵：

根据本发明的实施例，记观察序列概率矩阵为 $B = \{b_{i,j}(s_i)\}$ （ $i = 1,2,3,...,N$， $j \in Q$ ），其中 $b_{i,j}(s_i)$ 代表“在位点 i，考虑母亲单倍型和胎儿单倍型（状态 j， $j=\{x_i, y_i\}$ ）时，观测到这种测序信息的可能性”，即

$$b_{i,j}(s_i) = \Pr(s_i | q_i = j, \{m_0, m_1\})$$
$$= \frac{(n_{i,A} + n_{i,C} + n_{i,G} + n_{i,T})!}{n_{i,A}! n_{i,C}! n_{i,G}! n_{i,T}!} \cdot (P_{i,A})^{n_{i,A}} \cdot (P_{i,C})^{n_{i,C}} \cdot (P_{i,G})^{n_{i,G}} \cdot (P_{i,T})^{n_{i,T}},$$

其中$P_{i,base}$代表“在位点 i，考虑母亲单倍型和胎儿单倍型（状态 j，$j=\{x_i, y_i\}$）时，该碱基出现的可能性”，即

$$P_{i,base} = \Pr\left(base \middle| q_i = j, \{m_0, m_1\}\right) = \sum_{k\in\{0,1\}} \frac{1}{2}(1-\varepsilon)\Delta\left(base, m_k\right) + \frac{1}{2}\varepsilon\cdot\Delta\left(base, m_{x_i}\right) + \frac{1}{2}\varepsilon\cdot\Delta\left(base, f_{y_i}\right),$$

其中指示函数

$$\Delta(x,y) = \begin{cases} 1-e & x = y \\ e/3 & x \neq y \end{cases}$$

。

此步是进行 HMM 参数，每个位点的观察序列概率分布$b_{i,j}\left(s_i\right)$计算，即计算每个位点上不同胎儿单倍型（隐藏状态）下，血浆出现当前测序数据（观察序列）的可能性。

第三步，构建局部概率矩阵，和逆向指针（下面以一维局部概率矩阵构建为例）:

定义 局部概率 $\delta_i\left(q_i\right) = \left(\max_{q_{i-1}\in Q}\delta_i\left(q_i\right)\cdot a_{q_{i-1}q_i}\right)\cdot b_{i,q_i}\left(s_i\right)$

定义 逆向指针 $\Psi_i\left(q_i\right) = \arg\max_{q_{i-1}\in Q}\delta_i\left(q_i\right)\cdot a_{q_{i-1}q_i}$

这里所使用的术语局部概率$\delta_i\left(q_i\right)$和逆向指针$\Psi_i\left(q_i\right)$都是沿用 Viterbi 算法的经典定义。关于该参数的定义的详细描述，可以参见 Lawrence R. Rabiner, PROCEEDINGS OF THE IEEE, Vol.77，No.2，1989 年 2 月，通过参照将其全文并入本文。

第四步，确定最终状态，并回溯最优路径:

确定最终状态，$q_N^* = \arg\max_{q_N\in Q}\delta_N\left(q_N\right)$

按照逆向指针回溯最优路径，即最可能胎儿单基因型 $q_i^* = \Psi_i\left(q_i\right)$ $(i = 1,2,3,...,N-1)$。

第五步，输出结果。

由此，能够有效地对胎儿基因组的序列进行分析。相比其他已有的产前检测技术方法，本方法有以下技术优势，主要体现在准确性和可获得的遗传信息量上:

1) 根据本发明实施例，检测的位点不仅针对父源性位点，对母源性位点，即母亲杂合位点，同样可以很好地检测出来胎儿是否遗传了母亲的致病位点，检测准确率可高达 95%以上，且可以检测多种变异类型，扩大了疾病检测的范围。

2) 根据本发明实施例，不但可以通过一次测序获得多个位点、多种疾病的信息，对于一些在孕妇血浆中覆盖程度较低，单纯提高测序深度依然无法准确判定的基因序列，可以通过本方法推断得到，结果准确可靠。

3) 根据本发明实施例，可进行遗传疾病作图，对于一些连锁相关疾病，可通过其他位

点的信息直接推断出来，一次可获得的信息量大，对临床检测更加具有指导意义。

另外，根据本发明实施例，本发明的确定胎儿基因组中预定区域碱基信息的方法，不仅限于 SNP 或者 STR 等某一种遗传多态性位点，对所有的遗传多态性位点均可适用，且可以多种位点同时使用，以便互相验证。除了可进行产前无创检测胎儿基因组信息，达到疾病检测的目的，还可以进行无创产前亲子鉴定，在孩子出生前判定孩子父亲身份，为一些涉及抚养责任和义务、财产纠纷、性侵案等协助侦破。

用于确定胎儿基因组中预定区域碱基信息的系统

在本发明的又一方面，本发明提出了一种用于确定胎儿基因组中预定区域核酸序列的系统。根据本发明的实施例，参考图 2，该系统 1000 可以包括：文库构建装置 100、测序装置 200 以及分析装置 400。

根据本发明的实施例，文库构建装置 100 适于针对胎儿基因组 DNA 样本，构建测序文库。根据本发明的实施例，测序装置 200 与文库构建装置 100 相连，并且适于对所构建的测序文库进行测序，以便获得测序结果，所得到的测序结果由多个测序数据构成。根据本发明的实施例，还可以进一步包括 DNA 样本分离装置，该 DNA 样本分离装置适于从孕妇外周血中提取胎儿基因组 DNA 样本。由此，该系统可以适用于进行无创产前检测。

根据本发明的实施例，任选地，还可以包括比对装置 300。根据本发明的实施例，比对装置 300 与测序装置 200 相连，并且适于将所得到的测序结果与参照序列进行比对，以便确定与预定区域对应的测序数据。根据本发明的实施例，可以用于进行测序的方法和设备并不受特别限制，包括但不限于双脱氧链终止法；优选高通量的测序方法，由此，能够利用这些测序装置的高通量、深度测序的特点，进一步提高测序效率。从而，提高后续对测序数据进行分析，尤其是统计检验分析时的精确性和准确度。所述高通量的测序方法包括但不限于第二代测序技术或者是单分子测序技术。所述第二代测序平台（Metzker ML. Sequencing technologies-the next generation. Nat Rev Genet. 2010 Jan;11(1):31-46）包括但不限于 Illumina-Solexa（GA™,HiSeq2000™ 等）、ABI-Solid 和 Roche-454（焦磷酸测序）测序平台；单分子测序平台（技术）包括但不限于 Helicos 公司的真实单分子测序技术（True Single Molecule DNA sequencing），Pacific Biosciences 公司单分子实时测序（single molecule real-time (SMRT™)），以及 Oxford Nanopore Technologies 公司的纳米孔测序技术等（Rusk, Nicole (2009-04-01). Cheap Third-Generation Sequencing. Nature Methods 6 (4): 244-245）。随着测序技术的不断进化，本领域技术人员能够理解的是还可以采用其他的测序方法和装置进行全基因组测序。根据本发明的具体示例，可以利用选自 Illumina-Solexa、ABI-SOLiD、Roche-454 和单分子测序装置的至少一种对所述全基因组测序文库进行测序。根据本发明的实施例，可以使用的参照序列的类型并不受特别限制，可以为任何含有感兴趣区域的已知序列。根据本发明的实施例，可以采用已知的人类参考基因组作为参照序列。例如，根据本发明的实施例，采用的人类参考基因组为 NCBI 36.3，HG18。另外，根据本发明的实施例，进行比对的方法并不受特别限制。根据本发明的具体实施例，可以采用 SOAP 进行比

对。

根据本发明的实施例，分析装置 400 适于基于所述胎儿的测序结果，结合胎儿遗传相关个体的遗传信息，根据隐马尔可夫模型，确定所述预定区域的碱基信息。

根据本发明的实施例，惠特比算法采用 0.25 作为初始状态概率分布，采用 re/N 作为重组概率，其中 re=25~30，优选 25，N 为所述预定区域的长度，

采用

$$a_{jk}=\Pr\left(q_i=k|q_{i-1}=j\right)=\begin{cases}\left(1-p_r\right)^2 & x_i=x_{i-1},y_i=y_{i-1}\\ \left(1-p_r\right)\cdot p_r & x_i=x_{i-1},y_i\neq y_{i-1}\quad or\quad x_i\neq x_{i-1},y_i=y_{i-1}\\ p_r^2 & x_i\neq x_{i-1},y_i\neq y_{i-1}\end{cases}$$

作为重组转移矩阵， p_r=re/N。

根据本发明的实施例，将所述测序结果与参照序列进行比对，以便确定与所述预定区域对应的测序数据进一步包括按照下列公式确定概率最高的碱基:

$$P_{i,base}=\sum_{k\in\{0,1\}}\frac{1}{2}(1-\varepsilon)\Delta\left(base,m_k\right)+\frac{1}{2}\varepsilon\cdot\Delta\left(base,m_{x_i}\right)+\frac{1}{2}\varepsilon\cdot\Delta\left(base,f_{y_i}\right)$$

其中，

$$\Delta(x,y)=\begin{cases}1-e & x=y\\ e/3 & x\neq y\end{cases}$$

。

关于数据分析部分，前面已经进行了详细描述，也当然地适用于确定胎儿基因组中预定区域核酸序列的系统。不再赘述。

由此，利用该系统，能够有效地实施前面所述的确定胎儿基因组中预定区域核酸序列的方法，可以通过例如惠特比算法（Viterbi algorithm），借助隐马尔可夫模型，确定胎儿基因组中特定区域的碱基信息，由此，可以有效地对胎儿基因组的遗传信息进行产前检测。

此外，根据本发明的实施例，预定区域为已知存在遗传多态性的位点，而遗传多态性为选自单核苷酸多态性和 STR 的至少一种。

在本文中所述的术语“相连”应作广义理解，既可以是直接相连，也可以是间接相连，只要能够实现上述功能上的衔接即可。

需要说明的是，本领域技术人员能够理解，在前面所描述的确定胎儿基因组中预定区域核酸序列的方法的特征和优点也适合于确定胎儿基因组中预定区域核酸序列的系统，为描述方便，不再详述。

计算机可读介质

在本发明的又一方面，本发明提出了一种计算机可读介质。根据本发明的实施例，计

算机可读介质上存储有指令，所述指令适于被处理器执行以便基于胎儿的测序结果，结合胎儿遗传相关个体的遗传信息，根据隐马尔可夫模型，确定所述预定区域的碱基信息。由此，利用该计算机可读介质，能够有效地实施前面所述的方法，从而可以通过例如惠特比算法（Viterbi algorithm），借助隐马尔可夫模型，确定胎儿基因组中特定区域的碱基信息，由此，可以有效地对胎儿基因组的遗传信息进行产前检测。

根据本发明的实施例，指令适于按照惠特比算法，根据隐马尔可夫模型，确定所述预定区域的碱基信息。根据本发明的实施例，在所述惠特比算法中，采用 0.25 作为初始状态概率分布，采用 re/N 作为重组概率，其中 re=25~30，优选 25，N 为所述预定区域的长度，

采用

$$a_{jk}=\Pr\left(q_i=k\middle|q_{i-1}=j\right)=\begin{cases}\left(1-p_r\right)^2 & x_i=x_{i-1},y_i=y_{i-1}\\ \left(1-p_r\right)\cdot p_r & x_i=x_{i-1},y_i\neq y_{i-1}\quad or\quad x_i\neq x_{i-1},y_i=y_{i-1}\\ p_r^2 & x_i\neq x_{i-1},y_i\neq y_{i-1}\end{cases}$$

作为重组转移矩阵，其中，p_r=re/N。

根据本发明的实施例，所述指令将所述测序结果与参照序列进行比对，以便确定与所述预定区域对应的测序数据进一步包括按照下列公式确定概率最高的碱基:

$$P_{i,base}=\sum_{k\in\{0,1\}}\frac{1}{2}(1-\varepsilon)\Delta\left(base,m_k\right)+\frac{1}{2}\varepsilon\cdot\Delta\left(base,m_{x_i}\right)+\frac{1}{2}\varepsilon\cdot\Delta\left(base,f_{y_i}\right)$$

其中，

$$\Delta(x,y)=\begin{cases}1-e & x=y\\ e/3 & x\neq y\end{cases}$$

。

关于数据分析部分，前面已经进行了详细描述，也当然地适用于确定胎儿基因组中预定区域核酸序列的系统。不再赘述。

此外，根据本发明的实施例，预定区域为已知存在遗传多态性的位点，而遗传多态性为选自单核苷酸多态性和 STR 的至少一种。

就本说明书而言，“计算机可读介质”可以是任何可以包含、存储、通信、传播或传输程序以供指令执行系统、装置或设备或结合这些指令执行系统、装置或设备而使用的装置。计算机可读介质的更具体的示例（非穷尽性列表）包括以下：具有一个或多个布线的电连接部（电子装置），便携式计算机盘盒（磁装置），随机存取存储器（RAM），只读存储器（ROM），可擦除可编辑只读存储器（EPROM 或闪速存储器），光纤装置，以及便携式光盘只读存储器（CDROM）。另外，计算机可读介质甚至可以是可在其上打印所述程序的纸或其他合适的介质，因为例如可以通过对纸或其他介质进行光学扫描，接着进行编辑、解译或必要时以其他合适方式进行处理来以电子方式获得所述程序，然后将其存储在计算机

存储器中。

应当理解，本发明的各部分可以用硬件、软件、固件或它们的组合来实现。在上述实施方式中，多个步骤或方法可以用存储在存储器中且由合适的指令执行系统执行的软件或固件来实现。例如，如果用硬件来实现，和在另一实施方式中一样，可用本领域公知的下列技术中的任一项或他们的组合来实现：具有用于对数据信号实现逻辑功能的逻辑门电路的离散逻辑电路，具有合适的组合逻辑门电路的专用集成电路，可编程门阵列（PGA），现场可编程门阵列（FPGA）等。

本技术领域的普通技术人员可以理解实现上述实施例方法携带的全部或部分步骤是可以通过程序来指令相关的硬件完成，所述的程序可以存储于一种计算机可读存储介质中，该程序在执行时，包括方法实施例的步骤之一或其组合。

此外，在本发明各个实施例中的各功能单元可以集成在一个处理模块中，也可以是各个单元单独物理存在，也可以两个或两个以上单元集成在一个模块中。上述集成的模块既可以采用硬件的形式实现，也可以采用软件功能模块的形式实现。所述集成的模块如果以软件功能模块的形式实现并作为独立的产品销售或使用时，也可以存储在一个计算机可读取存储介质中。

下面将结合实施例对本发明的方案进行解释。本领域技术人员将会理解，下面的实施例仅用于说明本发明，而不应视为限定本发明的范围。实施例中未注明具体技术或条件的，按照本领域内的文献所描述的技术或条件（例如参考 J.萨姆布鲁克等著，黄培堂等译的《分子克隆实验指南》，第三版，科学出版社）或者按照产品说明书进行。所用试剂或仪器未注明生产厂商者，均为可以通过市购获得的常规产品，例如可以采购自 Illumina 公司。

一般方法

本发明实施例的主要步骤包括:

1) 无创采取含有胎儿遗传物质的孕妇样品，提取其中含有的遗传物质。
2) 胎儿家庭成员如父母和外祖父母等基因组 DNA 提取和纯化。
3) 各遗传物质根据不同测序平台测序要求进行文库构建。
4) 测序获得的数据进行过滤，过滤条件根据质量值、接头污染等来设定。
5) 获得的高质量序列根据需要进行组装处理，组装结果与人类基因组参考序列进行比对。获得唯一比对的序列，带入模型进行分析。

分析模型:

记号:

I.需要检测的位点数为 N。

II. 父母的单倍型分别记为 $FH=\{fh_0,fh_1\}$ 和 $MH=\{mh_0,mh_1\}$，

其中，

$mh_k = \{m_{1,k},...,m_{i,k},...,m_{N,k}\}$， $fh_k = \{f_{1,k},...,f_{i,k},...,f_{N,k}\}$，

$\forall fh_{i,k}, mh_{i,k} \in \{A,C,G,T\}$，

$k \in \{0,1\}$， $i = 1,2,3,...,N$。

III. 将未知的胎儿单倍型记为 $H = \{h_0, h_1\}$，特别地， h_0和 h_1分别遗传自母亲和父亲。

$h_0 = \{m_{1,x_1},...,m_{i,x_i},...,m_{N,x_N}\}$， $h_1 = \{f_{1,y_1},...,f_{i,y_i},...,f_{N,y_N}\}$，

其中， $x_i \in \{0,1\}$， $y_i \in \{0,1\}$，

下标 x_i和 y_i组成的序列对， $q_i = \{x_i, y_i\}$组成了我们需要解码的隐藏状态，

而所有可能出现的隐藏状态组成了集合 Q。

IV. 测序数据记为， $S = \{s_1,...,s_i,...,s_N\}$

其中， $s_i = \{n_{i,A}, n_{i,C}, n_{i,G}, n_{i,G}\}$，代表此位点的测序信息，包含了 ACGT 四种碱基的数量。

V. 平均胎儿浓度和平均测序错误率分别记为 ε和 e。

第一步，构建初始状态概率分布向量，以及单倍体重组转移矩阵：

I. 初始状态概率分布记为 $\pi = \{\pi_j\}$（ $j \in Q$），

根据本发明的实施例，在没有参考数据的情况下，可以设 $\pi_j = \Pr(q_1 = j) \triangleq \frac{1}{4}$，即每种隐藏状态在第一个位点出现的可能性相等。

II. 根据本发明的实施例， 记单倍体重组概率为 $p_r = re/N$，其中 re代表人类配子基因组重组平均次数，为先验数据在 25 到 30 之间。

III. 根据本发明的实施例，记单倍体重组转移矩阵记为 $A = \{a_{jk}\}$（ $j,k \in Q$），其中 a_{jk}为隐藏状态转移的概率，即

$$a_{jk} = \Pr(q_i = k | q_{i-1} = j) = \begin{cases} (1-p_r)^2 & x_i = x_{i-1}, y_i = y_{i-1} \\ (1-p_r)\cdot p_r & x_i = x_{i-1}, y_i \neq y_{i-1} \quad or \quad x_i \neq x_{i-1}, y_i = y_{i-1} \\ p_r^2 & x_i \neq x_{i-1}, y_i \neq y_{i-1} \end{cases}$$

第二步，构建观察序列概率矩阵：

根据本发明的实施例，记观察序列概率矩阵为 $B = \{b_{i,j}(s_i)\}$（ $i = 1,2,3,...,N$， $j \in Q$），其中 $b_{i,j}(s_i)$代表“在位点 i，考虑母亲单倍型和胎儿单倍型（状态 j）时，观测到这种测序

信息的可能性”，即

$$b_{i,j}(s_i) = \Pr\left(s_i \middle| q_i = j, \{m_0, m_1\}\right)$$
$$= \frac{(n_{i,A}+n_{i,C}+n_{i,G}+n_{i,T})!}{n_{i,A}!n_{i,C}!n_{i,G}!n_{i,T}!} \cdot \left(P_{i,A}\right)^{n_{i,A}} \cdot \left(P_{i,C}\right)^{n_{i,C}} \cdot \left(P_{i,G}\right)^{n_{i,G}} \cdot \left(P_{i,T}\right)^{n_{i,T}}$$

其中 $P_{i,base}$ 代表“在位点 i，考虑母亲单倍型和胎儿单倍型（状态 j）时，该碱基出现的可能性”，即

$$P_{i,base} = \Pr\left(base \middle| q_i = j, \{m_0, m_1\}\right)$$
$$= \sum_{k\in\{0,1\}} \frac{1}{2}(1-\varepsilon)\Delta\left(base, m_k\right) + \frac{1}{2}\varepsilon \cdot \Delta\left(base, m_{x_i}\right) + \frac{1}{2}\varepsilon \cdot \Delta\left(base, f_{y_i}\right),$$

其中指示函数

$$\Delta(x,y) = \begin{cases} 1-e & x = y \\ e/3 & x \neq y \end{cases}。$$

第三步，构建局部概率矩阵，和逆向指针（下面以一维局部概率矩阵构建为例）:

定义 局部概率 $\delta_i(q_i) = \left(\max_{q_{i-1}\in Q} \delta_i(q_i) \cdot a_{q_{i-1}q_i}\right) \cdot b_{i,q_i}(s_i)$

定义 逆向指针 $\Psi_i(q_i) = \arg\max_{q_{i-1}\in Q} \delta_i(q_i) \cdot a_{q_{i-1}q_i}$

第四步，确定最终状态状态，并回溯最优路径:

确定最终状态， $q_N^* = \arg\max_{q_N \in Q} \delta_N(q_N)$，

根据逆向指针回溯最优路径，即最可能胎儿单基因型 $q_i^* = \Psi_i(q_i)$ $(i = 1,2,3,...,N-1)$。

第五步，输出结果。

实施例 1

样品收集及处理:

（1）所取样品包括一个家庭中父亲和母亲孕期的外周血，胎儿出生后取脐带血，以EDTA 抗凝管收集，Oragene® DNA 唾液采集/DNA 纯化试剂盒 OG-250 采集祖父母和外祖父母唾液。

（2）祖父母和外祖父母唾液 DNA 提取后用 Infinium® HD Human610-Quad BeadChip 基因芯片进行分型。

（3）取母亲孕期外周血，1600g，4℃离心 10 分钟，将血细胞和血浆分开，血浆再以 16000g，4℃离心 10 分钟，进一步去除残留的白细胞。孕妇血浆用 TIANamp Micro DNA Kit（TIANGEN）提取 DNA，得到母亲和胎儿基因组 DNA 混合物，并且从全血分离血浆后剩余的白细胞中提取母亲基因组 DNA。将所得到的血浆 DNA 根据 Illumia®公司 HiSeq2000™

测序仪的上机要求进行建库，构建好的文库经 Agilent®Bioanalyzer 2100 检测片段分布范围符合要求，再经过 Q-PCR 方法对两个文库进行定量，合格后 Illumina® HiSeq2000™ 测序仪测序，测序循环数为 PE101index（即双向 101bp index 测序），其中仪器的参数设置及操作方法都按照 Illumina®操作手册（可由 http://www.illumina.com/support/documentation.ilmn 获取）。

（4）父亲外周血、母亲外周血白细胞和胎儿脐带血则直接用 TIANamp Micro DNA Kit（TIANGEN）提取试剂盒提取基因组 DNA。

除血浆 DNA 样品外，将所获得的所有 DNA 样品，需用 Covaris™ 打断仪打断至 500bp 大小的片段。将获得的 DNA 片段以及血浆 DNA 样品根据 Illumia®公司 HiSeq2000™ 测序仪的上机要求进行建库，具体步骤如下：

末端修复反应体系：

10 × T4 多核苷酸激酶缓冲液	10 μl
dNTPs(10mM)	4 μl
T4 DNA 聚合酶	5 μl
Klenow 片段	1 μl
T4 多核苷酸激酶	5 μl
DNA 片段	30μl
ddH_2O 补齐至 100 μl	

20℃反应 30 分钟后，使用 PCR Purification Kit(QIAGEN)回收末端修复产物。将所得到的产物最后溶于 34μl 的 EB 缓冲液中。

末端添加碱基 A 反应体系：

10 × Klenow 缓冲液	5μl
dATP(1mM)	10μl
Klenow (3'-5' exo^-)	3μl
DNA	32μl

37℃温育 30 分钟后，经 MinElute® PCR Purification Kit(QIAGEN)纯化并溶于 12μl 的 EB 中。

接头连接反应体系：

2x Rapid DNA 连接缓冲液	25μl
PEI Adapter oligomix(20uM)	10μl
T4 DNA 连接酶	5μl
添加碱基 A 的产物	10μl

20℃反应 15 分钟后，使用 PCR Purification Kit(QIAGEN)回收连接产物。将所得到的产物最后溶于 32μl 的 EB 缓冲液中。

PCR 反应体系：

接头连接反应产物	10 μl

Phusion DNA Polymerase Mix	25 μl
PCR 引物 (10 pmol/μl)	1 μl
Index N(10 pmol/μl)	1 μl
超纯水	13 μl

反应程序如下:

98℃	30 s	
98℃	10 s	10 个循环
65℃	30 s	
72℃	30 s	
72℃	5 min	
4℃	Hold	

使用 PCR Purification Kit(QIAGEN)回收 PCR 产物。样品最后溶于 50μl 的 EB 缓冲液中。

构建好的文库经 Agilent®Bioanalyzer 2100 检测片段分布范围符合要求，再经过 Q-PCR 方法对两个文库进行定量，合格后，用 Illumina® HiSeq2000™ 测序仪测序，测序循环数为 PE101index（即双向 101bp index 测序），其中仪器的参数设置及操作方法都按照 Illumina® 操作手册（可由 http://www.illumina.com/support/documentation.ilmn 获取）。

（5）父母基因组测序分型:

a. 使用 SOAP2 将测序数据比对到人类参考基因组（版本为 NCBI 36.3，HG18）。

b. 使用 SOAPsnp（南方汉族（CHS）家系数据使用的是千人计划数据）对数据进行一致序列（consensus sequence，CNS）构建。

c. 提取出标记位点的基因型。

（6）父母单倍体型推断:

a.构建含祖辈与父母基因型的群体基因型矩阵，即提取父母、祖辈和南方汉族家系在标记位点的基因型。

b. 使用 BEAGLE 对父母的单倍型进行推断。

（7）胎儿单倍体型推断:

a. 用 SOAP2 将血浆测序数据比对到人类参考基因组（版本为 NCBI 36.3，HG18）。

b. 构建初始状态概率向量，以及单倍体重组转移矩阵。

构建初始状态概率向量：采取无参考数据模式，即各个初始状态概率相等，均为 0.25。

单倍体重组转移矩阵：保守地，我们取 $re=25$（其余按一般方法所述）。

c. 统计每个位点的测序信息，并构建观察序列概率矩阵（其余按一般方法所述）。

d. 构建局部概率矩阵，和逆向指针（其余按一般方法所述）。

e. 确定最终状态状态，并回溯最优路径。

f. 输出。

根据胎儿出生后的脐带血基因分型结果，我们的分类准确性统计如下:

			母亲						合计		
			纯合			杂合					
			位点数	准确数	准确率	位点数	准确数	准确率	位点数	准确数	准确率
常染色体	父亲	纯合	199,552	199,552	100.00%	66,238	63,968	96.57%	265,790	263,520	99.15%
		杂合	65,409	64,735	98.97%	41,849	39,944	95.45%	107,258	104,679	97.60%
	合计		264,961	264,287	99.75%	108,087	103,912	96.14%	373,048	368,199	98.70%
X染色体			4,881	4,881	100.00%	1,718	1,478	86.03%	6,599	6,359	96.36%

工业实用性

本发明的确定胎儿基因组中预定区域碱基信息的方法、用于确定胎儿基因组中预定区域碱基信息的系统以及计算机可读介质，能够有效地应用于对胎儿基因组中预定区域的核酸序列进行分析。

尽管本发明的具体实施方式已经得到详细的描述，本领域技术人员将会理解。根据已经公开的所有教导，可以对那些细节进行各种修改和替换，这些改变均在本发明的保护范围之内。本发明的全部范围由所附权利要求及其任何等同物给出。

在本说明书的描述中，参考术语“一个实施例”、“一些实施例”、“示意性实施例”、“示例”、“具体示例”、或“一些示例”等的描述意指结合该实施例或示例描述的具体特征、结构、材料或者特点包含于本发明的至少一个实施例或示例中。在本说明书中，对上述术语的示意性表述不一定指的是相同的实施例或示例。而且，描述的具体特征、结构、材料或者特点可以在任何的一个或多个实施例或示例中以合适的方式结合。

权利要求书

1、一种确定胎儿基因组中预定区域碱基信息的方法，其特征在于，包括下列步骤：

针对胎儿基因组 DNA 样本，构建测序文库；

对所述测序文库进行测序，以便获得胎儿的测序结果，所述胎儿的测序结果由多个测序数据构成；以及

基于所述胎儿的测序结果，结合胎儿遗传相关个体的遗传信息，根据隐马尔可夫模型，确定所述预定区域的碱基信息。

2、根据权利要求 1 所述的方法，其特征在于，所述胎儿基因组 DNA 样本是从孕妇外周血中提取的。

3、根据权利要求 1 所述的方法，其特征在于，所述测序是利用选自 Illumina-Solexa、ABI-Solid、Roche-454 和单分子测序装置的至少一种对所述测序文库进行的。

4、根据权利要求 1 所述的方法，其特征在于，进一步包括将所述胎儿的测序结果与参照序列进行比对，以便确定来自于所述预定区域的测序结果。

5、根据权利要求 4 所述的方法，其特征在于，所述参照序列为人类参考基因组。

6、根据权利要求 1 所述的方法，其特征在于，所述胎儿遗传相关个体是所述胎儿的父母。

7、根据权利要求 1 所述的方法，其特征在于，按照惠特比算法，根据隐马尔可夫模型，确定所述预定区域的碱基信息。

8、根据权利要求 7 所述的方法，其特征在于，在所述惠特比算法中，采用 0.25 作为初始状态概率分布，采用 re/N 作为重组概率，其中 re=25~30，优选 25，N 为所述预定区域的长度，

采用

$$a_{jk}=\Pr\left(q_i=k\middle|q_{i-1}=j\right)=\begin{cases}\left(1-p_r\right)^2 & x_i=x_{i-1},y_i=y_{i-1}\\ \left(1-p_r\right)\cdot p_r & x_i=x_{i-1},y_i\neq y_{i-1}\quad or\quad x_i\neq x_{i-1},y_i=y_{i-1}\\ p_r^2 & x_i\neq x_{i-1},y_i\neq y_{i-1}\end{cases}$$

作为重组转移矩阵，其中，p_r=re/N。

9、根据权利要求 4 所述的方法，其特征在于，将所述胎儿的测序结果与参照序列进行比对，以便确定来自于所述预定区域的测序结果进一步包括按照下列公式确定概率最高的碱基：

$$P_{i,base}=\sum_{k\in\{0,1\}}\frac{1}{2}(1-\varepsilon)\Delta\left(base,m_k\right)+\frac{1}{2}\varepsilon\cdot\Delta\left(base,m_{x_i}\right)+\frac{1}{2}\varepsilon\cdot\Delta\left(base,f_{y_i}\right)$$

其中，

$$\Delta(x,y)=\begin{cases}1-e & x=y\\ e/3 & x\neq y\end{cases}$$

。

10、根据权利要求 1 所述的方法，其特征在于，进一步包括：所述预定区域为已知存在遗传多态性的位点。

11、根据权利要求 10 所述的方法，其特征在于，所述遗传多态性为选自单核苷酸多态性和 STR 的至少一种。

12、一种用于确定胎儿基因组中预定区域碱基信息的系统，其特征在于，包括：

文库构建装置，所述文库构建装置适于针对胎儿基因组 DNA 样本，构建测序文库；

测序装置，所述测序装置与所述文库构建装置相连，并且适于对所述测序文库进行测序，以便获得胎儿的测序结果，所述胎儿的测序结果由多个测序数据构成；

分析装置，基于所述胎儿的测序结果，结合胎儿遗传相关个体的遗传信息，根据隐马尔可夫模型，确定所述预定区域的碱基信息。

13、根据权利要求 12 所述的系统，其特征在于，进一步包括 DNA 样本分离装置，所述 DNA 样本分离装置适于从孕妇外周血中提取胎儿基因组 DNA 样本。

14、根据权利要求 12 所述的系统，其特征在于所述测序装置为选自 Illumina-Solexa、ABI-Solid、Roche-454 和单分子测序装置的至少一种。

15、根据权利要求 12 所述的系统，其特征在于，进一步包括比对装置，所述比对装置与所述测序装置相连，用于将所述胎儿的测序结果与参照序列进行比对，以便确定来自于所述预定区域的测序结果。

16、根据权利要求 12 所述的系统，其特征在于，所述分析装置适于按照惠特比算法，根据隐马尔可夫模型，确定所述预定区域的碱基信息。

17、根据权利要求 16 所述的系统，其特征在于，所述惠特比算法采用 0.25 作为初始状态概率分布，采用 re/N 作为重组概率，其中 re=25~30，优选 25，N 为所述预定区域的长度，

采用

$$a_{jk}=\Pr(q_i=k|q_{i-1}=j)=\begin{cases}(1-p_r)^2 & x_i=x_{i-1},y_i=y_{i-1}\\ (1-p_r)\cdot p_r & x_i=x_{i-1},y_i\neq y_{i-1}\quad or\quad x_i\neq x_{i-1},y_i=y_{i-1}\\ p_r^2 & x_i\neq x_{i-1},y_i\neq y_{i-1}\end{cases}$$

作为重组转移矩阵，其中，p_r=re/N。

18、根据权利要求 15 所述的系统，其特征在于，将所述胎儿的测序结果与参照序列进行比对，以便确定来自于所述预定区域的测序结果进一步包括按照下列公式确定概率最高的碱基：

$$P_{i,base} = \sum_{k\in\{0,1\}} \frac{1}{2}(1-\varepsilon)\Delta(base,m_k)+\frac{1}{2}\varepsilon\cdot\Delta(base,m_{x_i})+\frac{1}{2}\varepsilon\cdot\Delta(base,f_{y_i})$$

其中，

$$\Delta(x,y)=\begin{cases} 1-e & x=y \\ e/3 & x\neq y \end{cases}$$

。

19、一种计算机可读介质，其特征在于，所述计算机可读介质上存储有指令，所述指令适于被处理器执行以便基于胎儿的测序结果，结合胎儿遗传相关个体的遗传信息，根据隐马尔可夫模型，确定预定区域的碱基信息。

20、根据权利要求 19 所述的计算机可读介质，其特征在于，所述指令适于按照惠特比算法，根据隐马尔可夫模型，确定所述预定区域的碱基信息。

21、根据权利要求 20 所述的计算机可读介质，其特征在于，在所述惠特比算法中，采用 0.25 作为初始状态概率分布，采用 re/N 作为重组概率，其中 re=25~30，优选 25，N 为所述预定区域的长度，

采用

$$a_{jk}=\Pr(q_i=k|q_{i-1}=j)=\begin{cases} (1-p_r)^2 & x_i=x_{i-1},y_i=y_{i-1} \\ (1-p_r)\cdot p_r & x_i=x_{i-1},y_i\neq y_{i-1} \quad or \quad x_i\neq x_{i-1},y_i=y_{i-1} \\ p_r^2 & x_i\neq x_{i-1},y_i\neq y_{i-1} \end{cases}$$

作为重组转移矩阵，其中，p_r=re/N。

22、根据权利要求 19 所述的计算机可读介质，其特征在于，所述指令将所述胎儿的测序结果与参照序列进行比对，以便确定来自于所述预定区域的测序结果。

23、根据权利要求 22 所述的计算机可读介质，其特征在于，所述指令将所述胎儿的测序结果与参照序列进行比对，以便确定来自于所述预定区域的测序结果进一步包括按照下列公式确定概率最高的碱基:

$$P_{i,base} = \sum_{k\in\{0,1\}} \frac{1}{2}(1-\varepsilon)\Delta(base,m_k)+\frac{1}{2}\varepsilon\cdot\Delta(base,m_{x_i})+\frac{1}{2}\varepsilon\cdot\Delta(base,f_{y_i})$$

其中，

$$\Delta(x,y)=\begin{cases} 1-e & x=y \\ e/3 & x\neq y \end{cases}$$

。

24、根据权利要求 19 所述的计算机可读介质，其特征在于，进一步包括：所述预定区域为已知存在遗传多态性的位点。

25、根据权利要求 24 所述的计算机可读介质，其特征在于，所述遗传多态性为选自单核苷酸多态性和 STR 的至少一种。

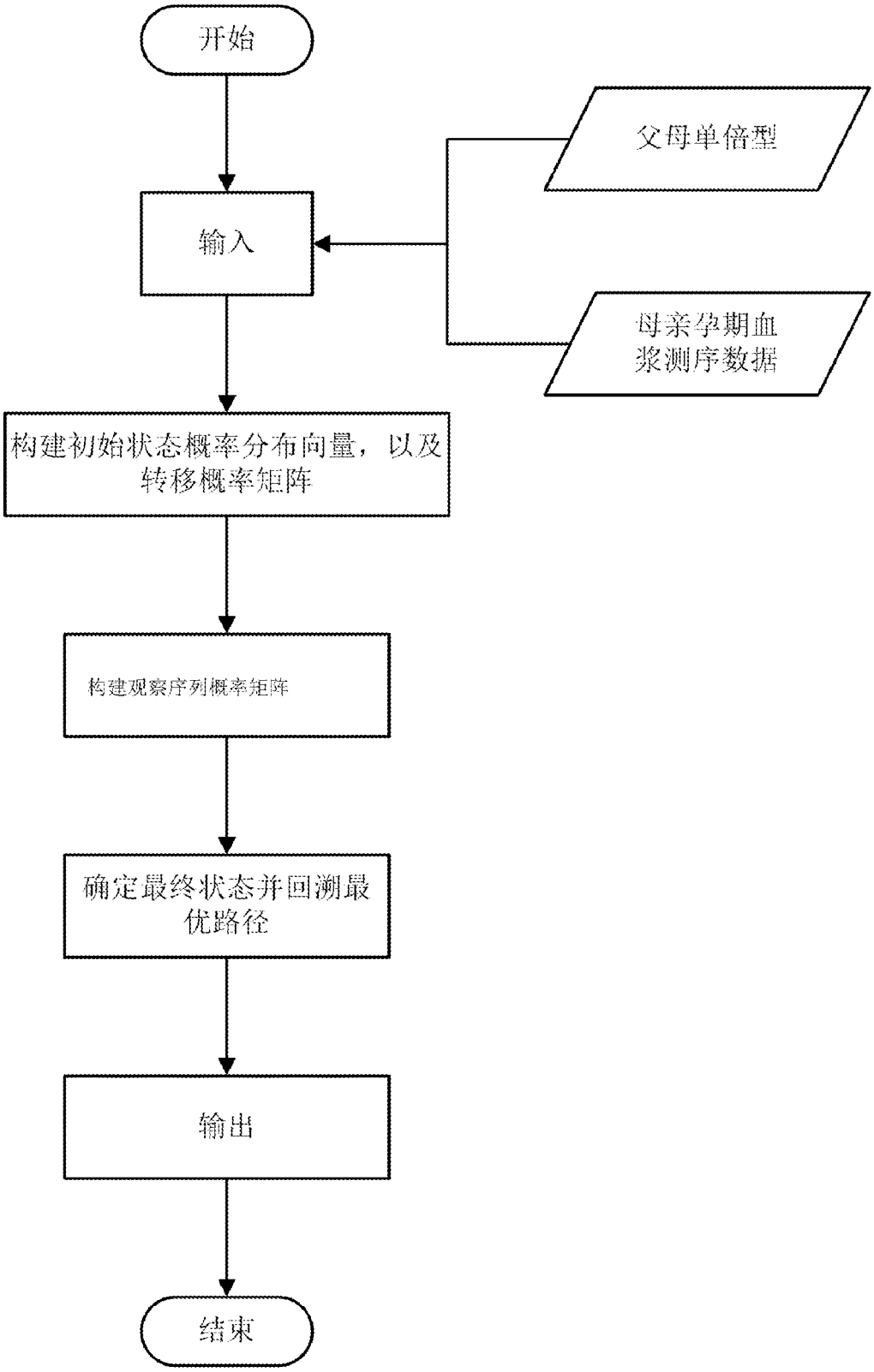

开始
父母单倍型
输入
母亲孕期血浆测序数据
构建初始状态概率分布向量，以及转移概率矩阵
构建观察序列概率矩阵
确定最终状态并回溯最优路径
输出
结束

图 1

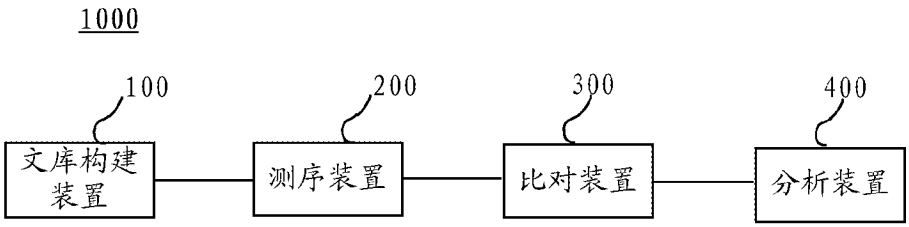


图 2

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2012/075478

A. CLASSIFICATION OF SUBJECT MATTER

C12Q 1/68 (2006.01) i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

IPC: C12Q

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

CNABS, CPRSABS, SIPOABS, DWPI, CNTXT, WOTXT, EPTXT, USTXT, CNKI, GOOGLE SCHOLAR, Elsevier Science, PubMed: hidden Markov model, HMM, viterbi algorithm, library, sequencing, fetus, foetus, fetal, foetal

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	CN 102127818 A (ZHANG, Kang) 20 July 2011 (20.07.2011) , claims 1-7, description, page 3, paragraph [0035] to page 4, paragraph [0050]	1-18
Y	CN 102317473 A (PACIFIC BIOSCIENCES CALIFORNIA INC.) 11 January 2012 (11.01.2012) , claims 1 and 2, description, page 32, paragraph [0003] to page 33, paragraph [0001]	1-18

☐ Further documents are listed in the continuation of Box C. ☒ See patent family annex.

* Special categories of cited documents:

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier application or patent but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&"document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
27 January 2013 (27.01.2013)	14 February 2013 (14.02.2013)

Name and mailing address of the ISA	Authorized officer
State Intellectual Property Office of the P. R. China No. 6, Xitucheng Road, Jimenqiao Haidian District, Beijing 100088, China Facsimile No. (86-10) 62019451	ZHAO, Yanhao Telephone No. (86-10) 62411043

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

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2012/075478

Box No. II Observations where certain claims were found unsearchable (Continuation of item 2 of first sheet)

This international search report has not been established in respect of certain claims under Article 17(2)(a) for the following reasons:

1. ☒ Claims Nos.:19-25
because they relate to subject matter not required to be searched by this Authority, namely:
Claims 19-25 are directed to mere presentations of information (PCT Rule 39. 1(v)).

2. ☐ Claims Nos.:
because they relate to parts of the international application that do not comply with the prescribed requirements to such an extent that no meaningful international search can be carried out, specifically:

3. ☐ Claims Nos.:
because they are dependent claims and are not drafted in accordance with the second and third sentences of Rule 6.4(a).

Box No. III Observations where unity of invention is lacking (Continuation of item 3 of first sheet)

This International Searching Authority found multiple inventions in this international application, as follows:

1. ☐ As all required additional search fees were timely paid by the applicant, this international search report covers all searchable claims.

2. ☐ As all searchable claims could be searched without effort justifying additional fees, this Authority did not invite payment of additional fees.

3. ☐ As only some of the required additional search fees were timely paid by the applicant, this international search report covers only those claims for which fees were paid, specifically claims Nos.:

4. ☐ No required additional search fees were timely paid by the applicant. Consequently, this international search report is restricted to the invention first mentioned in the claims; it is covered by claims Nos.:

Remark on protest

☐ The additional search fees were accompanied by the applicant's protest and, where applicable, the payment of a protest fee.

☐ The additional search fees were accompanied by the applicant's protest but the applicable protest fee was not paid within the time limit specified in the invitation.

☐ No protest accompanied the payment of additional search fees.

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2012/075478

Patent Documents referred in the Report	Publication Date	Patent Family	Publication Date
CN 102127818 A	20.07.2011	None	
CN 102317473 A	11.01.2012	WO 2010/068289 A2	17.06.2010
		US 2011/183320 A1	28.07.2011
		US 2010/221716 A1	02.09.2010
		CA 2746632 A1	17.06.2010
		WO 2010/068289 A3	14.10.2010
		AU 2009/325069 A1	17.06.2010
		EP 2370598 A2	05.10.2011

Form PCT/ISA /210 (patent family annex) (July 2009)

国际检索报告

国际申请号
PCT/CN2012/075478

A. 主题的分类

C12Q 1/68 (2006.01) i

按照国际专利分类(IPC)或者同时按照国家分类和 IPC 两种分类

B. 检索领域

检索的最低限度文献(标明分类系统和分类号)

IPC: C12Q

包含在检索领域中的除最低限度文献以外的检索文献

在国际检索时查阅的电子数据库(数据库的名称，和使用的检索词（如使用）)

CNABS,CPRSABS,SIPOABS,DWPI,CNTXT,WOTXT,EPTXT,USTXT,CNKI,GOOGLE SCHOLAR, Elsevier Science, PubMed：隐马尔可夫模型，惠特比算法，文库，测序，隐马可夫, 隐马氏，隐 MarkoV, 胎儿，hidden Markov model,HMM, viterbi algorithm, library, sequencing, fetus, foetus, fetal, foetal

C. 相关文件

类　型*	引用文件，必要时，指明相关段落	相关的权利要求
Y	CN102127818 A (张康) 20.7 月 2011（20.07.2011）,权利要求 1-7，说明书第 3 页第 35 段-说明书第 4 页第 50 段	1-18
Y	CN102317473 A (加利福尼亚太平洋生物科学股份有限公司) 11.1 月 2012（11.01.2012）,权利要求 1-2，说明书第 32 页第 3 段-第 33 页第 1 段	1-18

☐ 其余文件在 C 栏的续页中列出。　　☒ 见同族专利附件。

* 引用文件的具体类型:

"A" 认为不特别相关的表示了现有技术一般状态的文件

"E" 在国际申请日的当天或之后公布的在先申请或专利

"L" 可能对优先权要求构成怀疑的文件，或为确定另一篇引用文件的公布日而引用的或者因其他特殊理由而引用的文件（如具体说明的）

"O" 涉及口头公开、使用、展览或其他方式公开的文件

"P" 公布日先于国际申请日但迟于所要求的优先权日的文件

"T" 在申请日或优先权日之后公布，与申请不相抵触，但为了理解发明之理论或原理的在后文件

"X" 特别相关的文件，单独考虑该文件，认定要求保护的发明不是新颖的或不具有创造性

"Y" 特别相关的文件，当该文件与另一篇或者多篇该类文件结合并且这种结合对于本领域技术人员为显而易见时，要求保护的发明不具有创造性

"&" 同族专利的文件

国际检索实际完成的日期	国际检索报告邮寄日期
27.1 月 2013(27.01.2013)	**14.2 月 2013 (14.02.2013)**
ISA/CN 的名称和邮寄地址： 中华人民共和国国家知识产权局 中国北京市海淀区蓟门桥西土城路 6 号 100088 传真号：　(86-10)62019451	受权官员 赵彦豪 电话号码：(86-10) **62411043**

PCT/ISA/210 表(第 2 页) (2009 年 7 月)

国际检索报告

国际申请号
PCT/CN2012/075478

第II栏　某些权利要求被认为是不能检索的意见(续第 1 页第 2 项)

根据条约第 17 条(2)(a)，对某些权利要求未做国际检索报告的理由如下：

1. ☒ 权利要求：19-25

因为它们涉及不要求本单位进行检索的主题，即：

权利要求 19-25 涉及单纯的信息提供（PCT 细则 39.1 (v)）。

2. ☐ 权利要求：

因为它们涉及国际申请中不符合规定的要求的部分，以致不能进行任何有意义的国际检索，

具体地说：

3. ☐ 权利要求：

因为它们是从属权利要求，并且没有按照细则 6.4(a)第 2 句和第 3 句的要求撰写。

第III栏　缺乏发明单一性的意见(续第 1 页第 3 项)

本国际检索单位在该国际申请中发现多项发明，即：

1. ☐ 由于申请人按时缴纳了被要求缴纳的全部附加检索费，本国际检索报告涉及全部可作检索的权利要求。

2. ☐ 由于无需付出有理由要求附加费的劳动即能对全部可检索的权利要求进行检索，本单位未通知缴纳任何附加费。

3. ☐ 由于申请人仅按时缴纳了部分被要求缴纳的附加检索费，本国际检索报告仅涉及已缴费的那些权利要求。

具体地说，是权利要求：

4. ☐ 申请人未按时缴纳被要求缴纳的附加检索费。因此，本国际检索报告仅涉及权利要求书中首先提及的发明；包含该发明的权利要求是：

关于异议的说明： ☐ 申请人缴纳了附加检索费，同时提交了异议书，适用时，缴纳了异议费。

☐ 申请人缴纳了附加检索费，同时提交了异议书，但未在通知书规定的时间期限内缴纳异议费。

☐ 缴纳附加检索费时未提交异议书。

国际检索报告
关于同族专利的信息

国际申请号
PCT/CN2012/075478

检索报告中引用的专利文件	公布日期	同族专利	公布日期
CN102127818 A	20.07.2011	无	
CN 102317473 A	11.01.2012	WO 2010068289 A2	17.06.2010
		US 2011183320 A1	28.07.2011
		US 2010221716 A1	02.09.2010
		CA 2746632 A1	17.06.2010
		WO 2010068289 A3	14.10.2010
		AU 2009325069 A1	17.06.2010
		EP 2370598 A2	05.10.2011

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

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
19 December 2013 (19.12.2013)

WIPO | PCT

(10) International Publication Number
WO 2013/188846 A1

(51) International Patent Classification:
C12Q 1/68 (2006.01) *C40B 30/04* (2006.01)
G01N 33/53 (2006.01)

(21) International Application Number:
PCT/US2013/046020

(22) International Filing Date:
14 June 2013 (14.06.2013)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
61/660,427 15 June 2012 (15.06.2012) US

(72) Inventor; and
(71) Applicant : **STYLLI, Harry** [GB/US]; 2452 Paseo Dorado, La Jolla, CA 92037 (US).

(74) Agents: **HALEY, James, F.** et al.; Ropes & Gray LLP, 1211 Avenue of the Americas, New York, NY 10036 (US).

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

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

Published:

— *with international search report (Art. 21(3))*

— *before the expiration of the time limit for amending the claims and to be republished in the event of receipt of amendments (Rule 48.2(h))*

(54) Title: METHODS OF DETECTING DISEASES OR CONDITIONS

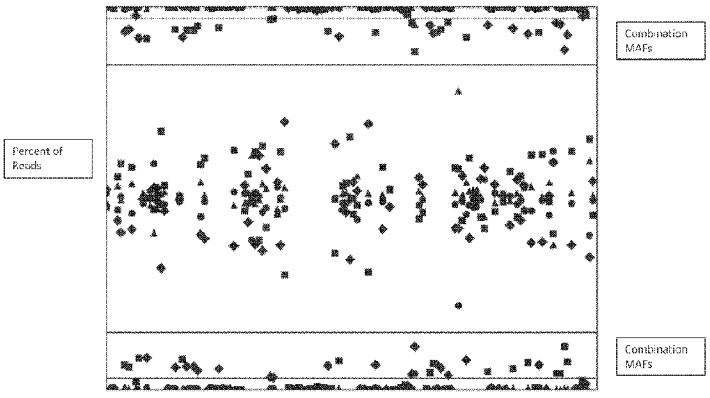


Figure 2

(57) Abstract: This invention provides methods of using a sample with multiple analytical components in the diagnosis, prognosis, or monitoring of diseases or conditions. The invention also provides methods of identifying markers of diseases or conditions.

WO 2013/188846 A1

METHODS OF DETECTING DISEASES OR CONDITIONS

Priority Information

[0001] This application claims priority from U.S. Provisional Application 61/660,427, filed June 15, 2012. The content and disclosure of that application is incorporated by reference herein in its entirety.

Field of the Invention

[0002] This invention relates generally to methods of using combinations of two or more different components selected from cell-free bodily fluids, phagocytic cells, circulating vesicles, and circulating diseased cells in the diagnosis, prognosis, or monitoring of a disease or condition. The invention also relates to methods of using the combinations to identify markers of diseases or conditions.

Background of the Invention

[0003] Leukocytes begin as pluripotent hematopoietic stem cells in the bone marrow and develop along either the myeloid lineage (monocytes, macrophages, neutrophils, eosinophils, and basophils) or the lymphoid lineage (T and B lymphocytes and natural killer cells). The major function of the myeloid lineage cells (e.g., neutrophils and macrophages) is the phagocytosis of infectious organisms, live unwanted damaged cells, senescent and dead cells (apoptotic and necrotic), as well as the clearing of cellular debris. Phagocytes from healthy animals do not replicate and are diploid, i.e., have a DNA content of 2n. On average, each cell contains <10 ng

DNA, <20 ng RNA, and <300 ng of protein. Non-phagocytic cells are also diploid and are not involved in the internalization of dead cells or infectious organisms and also have a DNA content of 2n.

[0004] The lifetime of various white blood cell subpopulations varies from a few days (e.g., neutrophils) to several months (e.g., macrophages). Like other cell types, leukocytes age and eventually die. During their aging process, human blood- and tissue-derived phagocytes (e.g., neutrophils) exhibit all the classic markers of programmed cell death (i.e., apoptosis), including caspase activation, pyknotic nuclei, and chromatin fragmentation. These cells also display a number of "eat-me" flags (e.g., phosphatidylserine, sugars) on the extracellular surfaces of their plasma membranes. Consequently, dying and dead cells and subcellular fragments thereof are cleared from tissues and blood by other phagocytic cells.

[0005] Early diagnosis of a disease often increases the likelihood of successful treatment or cure of such disease. Current diagnostic methods, however, focus on either using whole blood or separating the blood into different components, of which a single component is chosen for testing. Although this approach may enrich the signal being detected, it also results in the loss of potentially important information. Personalized diagnostic methods are needed that enable the diagnosis, especially the early diagnosis, of the presence of a disease or a condition in individuals who are not known to have the disease or who have recurrent disease.

[0006] One object of the present invention is to provide diagnostic methods that can facilitate the detection of a disease or condition-specific markers, e.g., nucleic acids, proteins, carbohydrates, and/or lipids and the like by using combinations of two or more different components selected from cell-free bodily fluids, phagocytic cells, circulating vesicles, and circulating diseased cells. Another object of this invention is to provide methods of identifying a disease or condition-specific markers and further use such markers alone or together with any known markers to diagnose diseases or conditions.

Summary of the Invention

[0007] Some embodiments of the invention are:

1. A method for diagnosing or aiding in the diagnosis of a disease or condition in a subject comprising:

a) determining a first profile of one or more markers of the disease or condition from a sample comprising two or more different components selected from the group consisting of: a cell-free bodily fluid isolated from the subject, a population of phagocytic cells isolated from the subject, a population of phagocytic cells having a DNA content more than 2n (>2n phagocytic cells) isolated from the subject, a population of circulating vesicles isolated from the subject, and a population of circulating diseased cells isolated from the subject;

b) determining a second profile of at least one of the one or more markers from a control comprising a component selected from the group consisting of: a population of phagocytic cells having a DNA content of 2n (=2n phagocytic cells) isolated from the subject, a population of non-phagocytic cells isolated from the subject, and a population of control cells isolated from the subject, wherein the control cells are substantially free of cells affected by the disease or condition; and

c) identifying a difference between the first and second profiles, wherein the difference is indicative of the presence of said disease or condition in the subject.

2. A method for assessing the risk of developing a disease or condition in a subject comprising:

a) determining a first profile of one or more markers of the disease or condition from a sample comprising two or more different components selected from the group consisting of: a cell-free bodily fluid isolated from the subject, a population of phagocytic cells isolated from the subject, a population of phagocytic cells having a DNA content more than 2n (>2n phagocytic cells) isolated from the subject, a population of circulating vesicles isolated from the subject, and a population of circulating diseased cells isolated from the subject;

b) determining a second profile of at least one of the one or more markers from a control comprising a component selected from the group consisting of: a population of phagocytic cells having a DNA content of 2n (=2n phagocytic cells) isolated from the subject, a population of non-phagocytic cells isolated from the subject, and a population of control cells isolated from the subject, wherein the control cells are substantially free of cells affected by the disease or condition; and

c) identifying a difference between the first and second profiles, wherein the difference is indicative of the risk of developing said disease or condition in the subject.

3. A method for prognosing or aiding in the prognosis of a disease or condition in a subject comprising:

a) determining a first profile of one or more markers of the disease or condition from a sample comprising two or more different components selected from the group consisting of: a cell-free bodily fluid isolated from the subject, a population of phagocytic cells isolated from the subject, a population of phagocytic cells having a DNA content more than 2n (>2n phagocytic cells) isolated from the subject, a population of circulating vesicles isolated from the subject, and a population of circulating diseased cells isolated from the subject;

b) determining a second profile of at least one of the one or more markers from a control comprising a component selected from the group consisting of: a population of phagocytic cells having a DNA content of 2n (=2n phagocytic cells) isolated from the subject, a population of non-phagocytic cells isolated from the subject, and a population of control cells isolated from the subject, wherein the control cells are substantially free of cells affected by the disease or condition; and

c) identifying a difference between the first and second profiles, wherein the difference is indicative of the prognosis of said disease or condition in the subject.

4. A method for assessing the efficacy of a treatment for a disease or condition in a subject comprising:

a) determining a first profile of one or more markers of the disease or condition from a first sample comprising two or more different components selected from the group consisting of: a cell-free bodily fluid isolated from the subject before the treatment, a population of phagocytic cells isolated from the subject before the treatment, a population of phagocytic cells having a DNA content more than 2n (>2n phagocytic cells) isolated from the subject before the treatment, a population of circulating vesicles isolated from the subject before the treatment, and a population of circulating diseased cells isolated from the subject before the treatment;

determining a second profile of at least one of the one or more markers from a first control comprising a component selected from the group consisting of: a population of phagocytic cells having a DNA content of 2n (=2n phagocytic cells) isolated from the subject before the treatment, a population of non-phagocytic cells isolated from the subject before the treatment, and a population of control cells isolated from the subject, wherein the control cells are substantially free of cells affected by the disease or condition, before the treatment;

identifying a difference between the first and second profiles;

b) determining a third profile of one or more markers of the disease or condition from a second sample comprising two or more different components selected from the group consisting of: a cell-free bodily fluid isolated from the subject after the treatment, a population of phagocytic cells isolated from the subject after the treatment, a population of phagocytic cells having a DNA content more than 2n (>2n phagocytic cells) isolated from the subject after the treatment, a population of circulating vesicles isolated from the subject after the treatment, and a population of circulating diseased cells isolated from the subject after the treatment;

determining a fourth profile of at least one of the one or more markers from a second control comprising a component selected from the group consisting of: a population of phagocytic cells having a DNA content of 2n (=2n phagocytic cells) isolated from the subject after the treatment, a population of non-phagocytic cells isolated from the subject after the treatment, and a population of control cells isolated

from the subject, wherein the control cells are substantially free of cells affected by the disease or condition, after the treatment;

identifying a difference between the third and fourth profiles; and

c) identifying a difference between the difference identified in a) and the difference identified in b) wherein the identified difference in c) is indicative of the efficacy of the treatment for said disease or condition in the subject.

5. A method for monitoring the progression or regression of a disease or condition in a subject comprising:

a) determining a first profile of one or more markers of the disease or condition from a first sample comprising two or more different components selected from the group consisting of: a cell-free bodily fluid isolated from the subject at a first time point, a population of phagocytic cells isolated from the subject at a first time point, a population of phagocytic cells having a DNA content more than 2n (>2n phagocytic cells) isolated from the subject at a first time point, a population of circulating vesicles isolated from the subject at a first time point, and a population of circulating diseased cells isolated from the subject at a first time point;

determining a second profile of at least one of the one or more markers from a first control comprising a component selected from the group consisting of: a population of phagocytic cells having a DNA content of 2n (=2n phagocytic cells) isolated from the subject at a first time point, a population of non-phagocytic cells isolated from the subject at a first time point, and a population of control cells isolated from the subject, wherein the control cells are substantially free of cells affected by the disease or condition, at a first time point;

identifying a difference between the first and second profiles;

b) determining a third profile of one or more markers of the disease or condition from a second sample comprising two or more different components selected from the group consisting of: a cell-free bodily fluid isolated from the subject at a second time point, a population of phagocytic cells isolated from the subject at a

second time point, a population of phagocytic cells having a DNA content more than 2n (>2n phagocytic cells) isolated from the subject at a second time point, a population of circulating vesicles isolated from the subject at a second time point, and a population of circulating diseased cells isolated from the subject at a second time point;

determining a fourth profile of at least one of the one or more markers from a second control comprising a component selected from the group consisting of: a population of phagocytic cells having a DNA content of 2n (=2n phagocytic cells) isolated from the subject at a second time point, a population of non-phagocytic cells isolated from the subject at a second time point, and a population of control cells isolated from the subject, wherein the control cells are substantially free of cells affected by the disease or condition, at a second time point;

identifying a difference between the third and fourth profiles; and

c) identifying a difference between the difference identified in a) and the difference identified in b) wherein the identified difference in c) is indicative of the progression or regression of said disease or condition in the subject.

6. A method for identifying a compound capable of ameliorating or treating a disease or condition in a subject comprising:

a) determining a first profile of one or more markers of the disease or condition from a first sample comprising two or more different components selected from the group consisting of: a cell-free bodily fluid isolated from the subject before administering the compound to the subject, a population of phagocytic cells isolated from the subject before administering the compound to the subject, a population of phagocytic cells having a DNA content more than 2n (>2n phagocytic cells) isolated from the subject before administering the compound to the subject, a population of circulating vesicles isolated from the subject before administering the compound to the subject, and a population of circulating diseased cells isolated from the subject before administering the compound to the subject;

determining a second profile of at least one of the one or more markers from a first control comprising a component selected from the group consisting of: a population of phagocytic cells having a DNA content of 2n (=2n phagocytic cells) isolated from the subject before administering the compound to the subject, a population of non-phagocytic cells isolated from the subject before administering the compound to the subject, and a population of control cells isolated from the subject, wherein the control cells are substantially free of cells affected by the disease or condition, before administering the compound to the subject;

identifying a difference between the first and second profiles;

b) determining a third profile of one or more markers of the disease or condition from a second sample comprising two or more different components selected from the group consisting of: a cell-free bodily fluid isolated from the subject after administering the compound to the subject, a population of phagocytic cells isolated from the subject after administering the compound to the subject, a population of phagocytic cells having a DNA content more than 2n (>2n phagocytic cells) isolated from the subject after administering the compound to the subject, a population of circulating vesicles isolated from the subject after administering the compound to the subject, and a population of circulating diseased cells isolated from the subject after administering the compound to the subject;

determining a fourth profile of at least one of the one or more markers from a second control comprising a component selected from the group consisting of: a population of phagocytic cells having a DNA content of 2n (=2n phagocytic cells) isolated from the subject after administering the compound to the subject, a population of non-phagocytic cells isolated from the subject after administering the compound to the subject, and a population of control cells isolated from the subject, wherein the control cells are substantially free of cells affected by the disease or condition, after administering the compound to the subject;

identifying a difference between the third and fourth profiles; and

c) identifying a difference between the difference identified in a) and the difference identified in b) wherein the identified difference in c) indicates that the compound is capable of ameliorating or treating said disease or condition in the subject.

7. A method for assessing the efficacy of a treatment for a disease or condition in a subject comprising:

a) determining a first profile of one or more markers of the disease or condition from a first sample comprising two or more different components selected from the group consisting of: a cell-free bodily fluid isolated from the subject before the treatment, a population of phagocytic cells isolated from the subject before the treatment, a population of phagocytic cells having a DNA content more than 2n (>2n phagocytic cells) isolated from the subject before the treatment, a population of circulating vesicles isolated from the subject before the treatment, and a population of circulating diseased cells isolated from the subject before the treatment;

b) determining a second profile of one or more markers of the disease or condition from a second sample comprising two or more different components selected from the group consisting of: a cell-free bodily fluid isolated from the subject after the treatment, a population of phagocytic cells isolated from the subject after the treatment, a population of phagocytic cells having a DNA content more than 2n (>2n phagocytic cells) isolated from the subject after the treatment, a population of circulating vesicles isolated from the subject after the treatment, and a population of circulating diseased cells isolated from the subject; and

c) identifying a difference between the first profile and the second profile, wherein the identified difference is indicative of the efficacy of the treatment for said disease or condition in the subject.

8. A method for monitoring the progression or regression of a disease or condition in a subject comprising:

a) determining a first profile of one or more markers of the disease or condition from a first sample comprising two or more different components selected

from the group consisting of: a cell-free bodily fluid isolated from the subject at a first time point, a population of phagocytic cells isolated from the subject at a first time point, a population of phagocytic cells having a DNA content more than 2n (>2n phagocytic cells) isolated from the subject at a first time point, a population of circulating vesicles isolated from the subject at a first time point, and a population of circulating diseased cells isolated from the subject at a first time point;

b) determining a second profile of one or more markers of the disease or condition from a second sample comprising two or more different components selected from the group consisting of: a cell-free bodily fluid isolated from the subject at a second time point, a population of phagocytic cells isolated from the subject at a second time point, a population of phagocytic cells having a DNA content more than 2n (>2n phagocytic cells) isolated from the subject at a second time point, a population of circulating vesicles isolated from the subject at a second time point, and a population of circulating diseased cells isolated from the subject at a second time point; and

c) identifying a difference between the first profile and the second profile, wherein the identified difference is indicative of the progression or regression of said disease or condition in the subject.

9. A method for identifying a compound capable of ameliorating or treating a disease or condition in a subject comprising:

a) determining a first profile of one or more markers of the disease or condition from a first sample comprising two or more different components selected from the group consisting of: a cell-free bodily fluid isolated from the subject before administering the compound to the subject, a population of phagocytic cells isolated from the subject before administering the compound to the subject, a population of phagocytic cells having a DNA content more than 2n (>2n phagocytic cells) isolated from the subject before administering the compound to the subject, a population of circulating vesicles isolated from the subject before administering the compound to the subject, and a population of circulating diseased cells isolated from the subject before administering the compound to the subject;

b) determining a second profile of one or more markers of the disease or condition from a second sample comprising two or more different components selected from the group consisting of: a cell-free bodily fluid isolated from the subject after administering the compound to the subject, a population of phagocytic cells isolated from the subject after administering the compound to the subject, a population of phagocytic cells having a DNA content more than 2n (>2n phagocytic cells) isolated from the subject after administering the compound to the subject, a population of circulating vesicles isolated from the subject after administering the compound to the subject, and a population of circulating diseased cells isolated from the subject after administering the compound to the subject; and

c) identifying a difference between the first profile and the second profile, wherein the identified difference indicates that the compound is capable of ameliorating or treating said disease or condition in the subject.

10. A method for diagnosing or aiding in the diagnosis of a disease or condition in a subject comprising:

a) determining a first profile of one or more markers of the disease or condition from a sample comprising components selected from the group consisting of: an analyte isolated from a cell-free bodily fluid isolated from the subject, an analyte isolated from a population of phagocytic cells isolated from the subject, an analyte isolated from a population of phagocytic cells having a DNA content more than 2n (>2n phagocytic cells) isolated from the subject, an analyte isolated from a population of circulating vesicles isolated from the subject, and an analyte isolated from a population of circulating diseased cells isolated from the subject;

b) determining a second profile of at least one of the one or more markers from a control comprising a component selected from the group consisting of: an analyte isolated from a population of phagocytic cells having a DNA content of 2n (=2n phagocytic cells) isolated from the subject, an analyte isolated from a population of non-phagocytic cells isolated from the subject, and an analyte isolated from a population of control cells isolated from the subject, wherein the control cells are substantially free of cells affected by the disease or condition; and

c) identifying a difference between the first and second profiles, wherein the difference is indicative of the presence of said disease or condition in the subject.

11. A method for assessing the risk of developing a disease or condition in a subject comprising:

a) determining a first profile of one or more markers of the disease or condition from a sample comprising two or more different components selected from the group consisting of: an analyte isolated from a cell-free bodily fluid isolated from the subject, an analyte isolated from a population of phagocytic cells isolated from the subject, an analyte isolated from a population of phagocytic cells having a DNA content more than 2n (>2n phagocytic cells) isolated from the subject, an analyte isolated from a population of circulating vesicles isolated from the subject, and an analyte isolated from a population of circulating diseased cells isolated from the subject;

b) determining a second profile of at least one of the one or more markers from a control comprising a component selected from the group consisting of: an analyte isolated from a population of phagocytic cells having a DNA content of 2n (=2n phagocytic cells) isolated from the subject, an analyte isolated from a population of non-phagocytic cells isolated from the subject, and an analyte isolated from a population of control cells isolated from the subject, wherein the control cells are substantially free of cells affected by the disease or condition; and

c) identifying a difference between the first and second profiles, wherein the difference is indicative of the risk of developing said disease or condition in the subject.

12. A method for prognosing or aiding in the prognosis of a disease or condition in a subject comprising:

a) determining a first profile of one or more markers of the disease or condition from a sample comprising two or more different components selected from the group consisting of: an analyte isolated from a cell-free bodily fluid isolated from the subject, an analyte isolated from a population of phagocytic cells isolated from the

subject, an analyte isolated from a population of phagocytic cells having a DNA content more than 2n (>2n phagocytic cells) isolated from the subject, an analyte isolated from a population of circulating vesicles isolated from the subject, and an analyte isolated from a population of circulating diseased cells isolated from the subject;

b) determining a second profile of at least one of the one or more markers from a control comprising a component selected from the group consisting of: an analyte isolated from a population of phagocytic cells having a DNA content of 2n (=2n phagocytic cells) isolated from the subject, an analyte isolated from a population of non-phagocytic cells isolated from the subject, and an analyte isolated from a population of control cells isolated from the subject, wherein the control cells are substantially free of cells affected by the disease or condition; and

c) identifying a difference between the first and second profiles, wherein the difference is indicative of the prognosis of said disease or condition in the subject.

13. A method for assessing the efficacy of a treatment for a disease or condition in a subject comprising:

a) determining a first profile of one or more markers of the disease or condition from a first sample comprising two or more different components selected from the group consisting of: an analyte isolated from a cell-free bodily fluid isolated from the subject before the treatment, an analyte isolated from a population of phagocytic cells isolated from the subject before the treatment, an analyte isolated from a population of phagocytic cells having a DNA content more than 2n (>2n phagocytic cells) isolated from the subject before the treatment, an analyte isolated from a population of circulating vesicles isolated from the subject before the treatment, and an analyte isolated from a population of circulating diseased cells isolated from the subject before the treatment;

determining a second profile of at least one of the one or more markers from a first control comprising a component selected from the group consisting of: an analyte isolated from a population of phagocytic cells having a DNA content of 2n

(=2n phagocytic cells) isolated from the subject before the treatment, an analyte isolated from a population of non-phagocytic cells isolated from the subject before the treatment, and an analyte isolated from a population of control cells isolated from the subject, wherein the control cells are substantially free of cells affected by the disease or condition, before the treatment;

identifying a difference between the first and second profiles;

b) determining a third profile of one or more markers of the disease or condition from a second sample comprising two or more different components selected from the group consisting of: an analyte isolated from a cell-free bodily fluid isolated from the subject after the treatment, an analyte isolated from a population of phagocytic cells isolated from the subject after the treatment, an analyte isolated from a population of phagocytic cells having a DNA content more than 2n (>2n phagocytic cells) isolated from the subject after the treatment, an analyte isolated from a population of circulating vesicles isolated from the subject after the treatment, and an analyte isolated from a population of circulating diseased cells isolated from the subject after the treatment;

determining a fourth profile of at least one of the one or more markers from a second control comprising a component selected from the group consisting of: an analyte isolated from a population of phagocytic cells having a DNA content of 2n (=2n phagocytic cells) isolated from the subject after the treatment, an analyte isolated from a population of non-phagocytic cells isolated from the subject after the treatment, and an analyte isolated from a population of control cells isolated from the subject, wherein the control cells are substantially free of cells affected by the disease or condition, after the treatment;

identifying a difference between the third and fourth profiles; and

c) identifying a difference between the difference identified in a) and the difference identified in b) wherein the identified difference in c) is indicative of the efficacy of the treatment for said disease or condition in the subject.

14. A method for monitoring the progression or regression of a disease or condition in a subject comprising:

a) determining a first profile of one or more markers of the disease or condition from a first sample comprising two or more different components selected from the group consisting of: an analyte isolated from a cell-free bodily fluid isolated from the subject at a first time point, an analyte isolated from a population of phagocytic cells isolated from the subject at a first time point, an analyte isolated from a population of phagocytic cells having a DNA content more than 2n (>2n phagocytic cells) isolated from the subject at a first time point, an analyte isolated from a population of circulating vesicles isolated from the subject at a first time point, and an analyte isolated from a population of circulating diseased cells isolated from the subject at a first time point;

determining a second profile of at least one of the one or more markers from a first control comprising a component selected from the group consisting of: an analyte isolated from a population of phagocytic cells having a DNA content of 2n (=2n phagocytic cells) isolated from the subject at a first time point, an analyte isolated from a population of non-phagocytic cells isolated from the subject at a first time point, and an analyte isolated from a population of control cells isolated from the subject, wherein the control cells are substantially free of cells affected by the disease or condition, at a first time point;

identifying a difference between the first and second profiles;

b) determining a third profile of one or more markers of the disease or condition from a second sample comprising two or more different components selected from the group consisting of: an analyte isolated from a cell-free bodily fluid isolated from the subject at a second time point, an analyte isolated from a population of phagocytic cells isolated from the subject at a second time point, an analyte isolated from a population of phagocytic cells having a DNA content more than 2n (>2n phagocytic cells) isolated from the subject at a second time point, an analyte isolated from a population of circulating vesicles isolated from the subject at a second

time point, and an analyte isolated from a population of circulating diseased cells isolated from the subject at a second time point;

determining a fourth profile of at least one of the one or more markers from a second control comprising a component selected from the group consisting of: an analyte isolated from a population of phagocytic cells having a DNA content of 2n (=2n phagocytic cells) isolated from the subject at a second time point, an analyte isolated from a population of non-phagocytic cells isolated from the subject at a second time point, and an analyte isolated from a population of control cells isolated from the subject, wherein the control cells are substantially free of cells affected by the disease or condition, at a second time point;

identifying a difference between the third and fourth profiles; and

c) identifying a difference between the difference identified in a) and the difference identified in b) wherein the identified difference in c) is indicative of the progression or regression of said disease or condition in the subject.

15. A method for identifying a compound capable of ameliorating or treating a disease or condition in a subject comprising:

a) determining a first profile of one or more markers of the disease or condition from a first sample comprising two or more different components selected from the group consisting of: an analyte isolated from a cell-free bodily fluid isolated from the subject before administering the compound to the subject, an analyte isolated from a population of phagocytic cells isolated from the subject before administering the compound to the subject, an analyte isolated from a population of phagocytic cells having a DNA content more than 2n (>2n phagocytic cells) isolated from the subject before administering the compound to the subject, an analyte isolated from a population of circulating vesicles isolated from the subject before administering the compound to the subject, and an analyte isolated from a population of circulating diseased cells isolated from the subject before administering the compound to the subject;

determining a second profile of at least one of the one or more markers from a first control comprising a component selected from the group consisting of: an analyte isolated from a population of phagocytic cells having a DNA content of 2n (=2n phagocytic cells) isolated from the subject before administering the compound to the subject, an analyte isolated from a population of non-phagocytic cells isolated from the subject before administering the compound to the subject, and an analyte isolated from a population of control cells isolated from the subject, wherein the control cells are substantially free of cells affected by the disease or condition, before administering the compound to the subject;

identifying a difference between the first and second profiles;

b) determining a third profile of one or more markers of the disease or condition from a second sample comprising two or more different components selected from the group consisting of: an analyte isolated from a cell-free bodily fluid isolated from the subject after administering the compound to the subject, an analyte isolated from a population of phagocytic cells isolated from the subject after administering the compound to the subject, an analyte isolated from a population of phagocytic cells having a DNA content more than 2n (>2n phagocytic cells) isolated from the subject after administering the compound to the subject, an analyte isolated from a population of circulating vesicles isolated from the subject after administering the compound to the subject, and an analyte isolated from a population of circulating diseased cells isolated from the subject after administering the compound to the subject;

determining a fourth profile of at least one of the one or more markers from a second control comprising a component selected from the group consisting of: an analyte isolated from a population of phagocytic cells having a DNA content of 2n (=2n phagocytic cells) isolated from the subject after administering the compound to the subject, an analyte isolated from a population of non-phagocytic cells isolated from the subject after administering the compound to the subject, and an analyte isolated from a population of control cells isolated from the subject, wherein the

control cells are substantially free of cells affected by the disease or condition, after administering the compound to the subject;

identifying a difference between the third and fourth profiles; and

c) identifying a difference between the difference identified in a) and the difference identified in b) wherein the identified difference in c) indicates that the compound is capable of ameliorating or treating said disease or condition in the subject.

16. A method for assessing the efficacy of a treatment for a disease or condition in a subject comprising:

a) determining a first profile of one or more markers of the disease or condition from a first sample comprising two or more different components selected from the group consisting of: an analyte isolated from a cell-free bodily fluid isolated from the subject before the treatment, a population of phagocytic cells isolated from the subject before the treatment, an analyte isolated from a population of phagocytic cells having a DNA content more than 2n (>2n phagocytic cells) isolated from the subject before the treatment, an analyte isolated from a population of circulating vesicles isolated from the subject before the treatment, and an analyte isolated from a population of circulating diseased cells isolated from the subject before the treatment;

b) determining a second profile of one or more markers of the disease or condition from a second sample comprising two or more different components selected from the group consisting of: an analyte isolated from a cell-free bodily fluid isolated from the subject after the treatment, an analyte isolated from a population of phagocytic cells isolated from the subject after the treatment, an analyte isolated from a population of phagocytic cells having a DNA content more than 2n (>2n phagocytic cells) isolated from the subject after the treatment, an analyte isolated from a population of circulating vesicles isolated from the subject after the treatment, and an analyte isolated from a population of control cells isolated from the subject, wherein the control cells are substantially free of cells affected by the disease or condition; and

c) identifying a difference between the first profile and the second profile, wherein the identified difference is indicative of the efficacy of the treatment for said disease or condition in the subject.

17. A method for monitoring the progression or regression of a disease or condition in a subject comprising:

a) determining a first profile of one or more markers of the disease or condition from a first sample comprising two or more different components selected from the group consisting of: an analyte isolated from a cell-free bodily fluid isolated from the subject at a first time point, an analyte isolated from a population of phagocytic cells isolated from the subject at a first time point, an analyte isolated from a population of phagocytic cells having a DNA content more than 2n (>2n phagocytic cells) isolated from the subject at a first time point, an analyte isolated from a population of circulating vesicles isolated from the subject at a first time point, and an analyte isolated from a population of circulating diseased cells isolated from the subject at a first time point;

b) determining a second profile of one or more markers of the disease or condition from a second sample comprising two or more different components selected from the group consisting of: an analyte isolated from a cell-free bodily fluid isolated from the subject at a second time point, an analyte isolated from a population of phagocytic cells isolated from the subject at a second time point, an analyte isolated from a population of phagocytic cells having a DNA content more than 2n (>2n phagocytic cells) isolated from the subject at a second time point, an analyte isolated from a population of circulating vesicles isolated from the subject at a second time point, and an analyte isolated from a population of circulating diseased cells isolated from the subject at a second time point; and

c) identifying a difference between the first profile and the second profile, wherein the identified difference is indicative of the progression or regression of said disease or condition in the subject.

18. A method for identifying a compound capable of ameliorating or treating a disease or condition in a subject comprising:

a) determining a first profile of one or more markers of the disease or condition from a first sample comprising two or more different components selected from the group consisting of: an analyte isolated from a cell-free bodily fluid isolated from the subject before administering the compound to the subject, an analyte isolated from a population of phagocytic cells isolated from the subject before administering the compound to the subject, an analyte isolated from a population of phagocytic cells having a DNA content more than 2n (>2n phagocytic cells) isolated from the subject before administering the compound to the subject, an analyte isolated from a population of circulating vesicles isolated from the subject before administering the compound to the subject, and an analyte isolated from a population of circulating diseased cells isolated from the subject before administering the compound to the subject;

b) determining a second profile of one or more markers of the disease or condition from a second sample comprising two or more different components selected from the group consisting of: an analyte isolated from a cell-free bodily fluid isolated from the subject after administering the compound to the subject, an analyte isolated from a population of phagocytic cells isolated from the subject after administering the compound to the subject, an analyte isolated from a population of phagocytic cells having a DNA content more than 2n (>2n phagocytic cells) isolated from the subject after administering the compound to the subject, an analyte isolated from a population of circulating vesicles isolated from the subject after administering the compound to the subject, and an analyte isolated from a population of circulating diseased cells isolated from the subject after administering the compound to the subject; and

c) identifying a difference between the first profile and the second profile, wherein the identified difference indicates that the compound is capable of ameliorating or treating said disease or condition in the subject.

19. A method for diagnosing or aiding in the diagnosis of a disease or condition in a subject comprising:

a) determining a first profile of one or more markers of the disease or condition from a sample comprising two or more different components selected from the group consisting of: a cell-free bodily fluid isolated from the subject, a population of phagocytic cells isolated from the subject, a population of phagocytic cells having a DNA content more than 2n (>2n phagocytic cells) isolated from the subject, a population of circulating vesicles isolated from the subject, and a population of circulating diseased cells isolated from the subject; and

b) identifying a difference between the first profile and a second profile of at least one of the one or more markers from a repository of said markers of said disease or condition, wherein the difference is indicative of the presence of said disease or condition in the subject.

20. A method for assessing the risk of developing a disease or condition in a subject comprising:

a) determining a first profile of one or more markers of the disease or condition from a sample comprising two or more different components selected from the group consisting of: a cell-free bodily fluid isolated from the subject, a population of phagocytic cells isolated from the subject, a population of phagocytic cells having a DNA content more than 2n (>2n phagocytic cells) isolated from the subject, a population of circulating vesicles isolated from the subject, and a population of circulating diseased cells isolated from the subject; and

b) identifying a difference between the first profile and a second profile of at least one of the one or more markers from a repository of said markers of said disease or condition , wherein the difference is indicative of the risk of developing said disease or condition in the subject.

21. A method for prognosing or aiding in the prognosis of a disease or condition in a subject comprising:

a) determining a first profile of one or more markers of the disease or condition from a sample comprising two or more different components selected from the group consisting of: a cell-free bodily fluid isolated from the subject, a population of phagocytic cells isolated from the subject, a population of phagocytic cells having a DNA content more than 2n (>2n phagocytic cells) isolated from the subject, a population of circulating vesicles isolated from the subject, and a population of circulating diseased cells isolated from the subject; and

b) identifying a difference between the first profile and a second profile of at least one of the one or more markers from a repository of said markers of said disease or condition, wherein the difference is indicative of the prognosis of said disease or condition in the subject.

22. A method for diagnosing or aiding in the diagnosis of a disease or condition in a subject comprising:

a) determining a first profile of one or more markers of the disease or condition from a sample comprising components selected from the group consisting of: an analyte isolated from a cell-free bodily fluid isolated from the subject, an analyte isolated from a population of phagocytic cells isolated from the subject, an analyte isolated from a population of phagocytic cells having a DNA content more than 2n (>2n phagocytic cells) isolated from the subject, an analyte isolated from a population of circulating vesicles isolated from the subject, and an analyte isolated from a population of circulating diseased cells isolated from the subject; and

b) identifying a difference between the first profile and a second profile of at least one of the one or more markers from a repository of said markers of said disease or condition , wherein the difference is indicative of the presence of said disease or condition in the subject.

23. A method for assessing the risk of developing a disease or condition in a subject comprising:

a) determining a first profile of one or more markers of the disease or condition from a sample comprising two or more different components selected from

the group consisting of: an analyte isolated from a cell-free bodily fluid isolated from the subject, an analyte isolated from a population of phagocytic cells isolated from the subject, an analyte isolated from a population of phagocytic cells having a DNA content more than 2n (>2n phagocytic cells) isolated from the subject, an analyte isolated from a population of circulating vesicles isolated from the subject, and an analyte isolated from a population of circulating diseased cells isolated from the subject; and

b) identifying a difference between the first profile and a second profile of at least one of the one or more markers from a repository of said markers of said disease or condition , wherein the difference is indicative of the risk of developing said disease or condition in the subject.

24. A method for prognosing or aiding in the prognosis of a disease or condition in a subject comprising:

a) determining a first profile of one or more markers of the disease or condition from a sample comprising two or more different components selected from the group consisting of: an analyte isolated from a cell-free bodily fluid isolated from the subject, an analyte isolated from a population of phagocytic cells isolated from the subject, an analyte isolated from a population of phagocytic cells having a DNA content more than 2n (>2n phagocytic cells) isolated from the subject, an analyte isolated from a population of circulating vesicles isolated from the subject, and a population of circulating diseased cells isolated from the subject; and

b) identifying a difference between the first profile and a second profile of at least one of the one or more markers from a repository of said markers of said disease or condition , wherein the difference is indicative of the prognosis of said disease or condition in the subject.

25. The method of any one of embodiments 1-3 and 10-12, wherein at least one of the one or more markers is up-regulated or activated in the sample compared to the control.

26. The method of any one of embodiments 1-3 and 10-12, wherein at least one of the one or more markers is down-regulated or inhibited in the sample compared to the control.

27. The method of any one of embodiments 4-6 and 13-15, wherein at least one of the one or more markers is up-regulated or activated in the first sample compared to the first control or in the second sample compared to the second control.

28. The method of any one of embodiments 4-6 and 13-15, wherein at least one of the one or more markers is down-regulated or inhibited in the first sample compared to the first control or in the second sample compared to the second control.

29. The method of any one of embodiments 7-9 and 16-18, wherein at least one of the one or more markers is up-regulated or activated in the first sample compared to the second sample.

30. The method of any one of embodiments 7-9 and 16-18, wherein at least one of the one or more markers is down-regulated or inhibited in the first sample compared to the second sample.

31. The method of any one of embodiments 19-24, wherein at least one of the one or more markers is up-regulated or activated in the sample compared to the repository.

32. The method of any one of embodiments 19-24, wherein at least one of the one or more markers is down-regulated or inhibited in the sample compared to the repository.

33. The method of any one of embodiments 1-32, wherein the first profile or the second profile comprises the absence of at least one of the one or more markers of said disease or condition.

34. The method of any one of embodiments 4-6 and 13-15, wherein the third profile or the fourth profile comprises the absence of at least one of the one or more markers of said disease or condition.

35. The method of any one of embodiments 1-9 and 19-21, wherein, when said method comprises circulating diseased cells, control cells not affected by the disease or condition, phagocytic cells, >2n phagocytic cells, =2n phagocytic cells, or non-phagocytic cells, the method further comprises lysing the circulating diseased cells, control cells not affected by the disease or condition, phagocytic cells, >2n phagocytic cells, =2n phagocytic cells, or non-phagocytic cells.

36. The method of any one of embodiments 1-9, 19-21, and 35 wherein, when said method comprises circulating diseased cells, control cells not affected by the disease or condition, phagocytic cells, >2n phagocytic cells, =2n phagocytic cells, or non-phagocytic cells, the method further comprises extracting at least some of the cellular contents from the circulating diseased cells, control cells not affected by the disease or condition, phagocytic cells, >2n phagocytic cells, =2n phagocytic cells, or non-phagocytic cells.

37. The method of any one of embodiments 1-9 and 19-21, 35, and 36, wherein, when said method comprises the cell-free bodily fluid, the method further comprises extracting the one or more markers from the cell-free bodily fluid.

38. The method of any one of embodiments 1-9 and 19-21, wherein, when said method comprises the cell-free bodily fluid, the cell-free bodily fluid comprises a transrenal nucleic acid.

39. The method of any one of embodiments 1-38, wherein at least one of the one or more markers of said disease or condition is present in the circulating diseased cells, cell-free bodily fluid sample, phagocytic cells, or >2n phagocytic cells.

40. The method of any one of embodiments 1-39, wherein at least one of the one or more markers of said disease or condition is not present in the circulating diseased cells, cell-free bodily fluid sample, phagocytic cells, or >2n phagocytic cells.

41. The method of any one of embodiments 1-40, wherein one or more of the circulating diseased cells, control cells not affected by the disease or condition, phagocytic cells, >2n phagocytic cells, =2n phagocytic cells, or non-phagocytic cells are enucleated.

42. The method of embodiment 41, wherein the cells are enucleated using physical removal, chemical treatments, photoablation, or ultraviolet irradiation.

43. The method of embodiment 42, wherein the physical removal uses a microneedle, optical tweezer or aspiration.

44. The method of any one of embodiments 19-24, wherein the repository is obtained by data mining.

45. The method of any one of embodiments 1-44, wherein the phagocytic cells, >2n phagocytic cells, or =2n phagocytic cells are neutrophils, macrophages, monocytes, dendritic cells, foam cells, mast cells, eosinophils, keratinocytes, or mixtures thereof.

46. The method of any one of embodiments 1-45, wherein the non-phagocytic cells are T cells, B cells, null cells, basophils, or mixtures thereof.

47. The method of any one of embodiments 1-46, wherein the circulating diseased cells are blood cells, tumor cells, lymphoma cells, fetal cells, apoptotic cells, epithelia cells, endothelial cells, stem cells, progenitor cells, mesenchymal cells, osteoblast cells, osteocytes, hematopoietic stem cells, foam cells, adipose cells, transcervical cells, circulating cardiocytes, circulating fibrocytes, circulating myocytes, circulating cells from kidney, circulating cells from gastrointestinal tract, circulating cells from lung, circulating cells from reproductive organs, circulating cells from central nervous system, circulating hepatic cells, circulating cells from spleen, circulating cells from thymus, circulating cells from thyroid , circulating cells from an endocrine gland, circulating cells from parathyroid, circulating cells from pituitary, circulating cells from adrenal gland, circulating cells from islets of Langerhans, circulating cells from pancreas, circulating cells from hypothalamus, circulating cells from prostate tissues, circulating cells from breast tissues, circulating cells from circulating retinal cells, circulating ophthalmic cells, circulating auditory cells, circulating epidermal cells, circulating cells from the urinary tract, or mixtures thereof.

48. The method of any one of embodiments 1-47, wherein the control cells are normal cells.

49. The method of any one of embodiments 1-48, wherein the control cells are circulating cells.

50. The method of any one of embodiments 1-49, wherein the circulating vesicles are selected from the group consisting of circulating microvesicles, apoptotic bodies, micro-particles, membrane-bound vesicles, multivesicular bodies, nanovesicles, microparticles, and ARRDC-1 mediated microvesicles (ARMM).

51. The method of embodiment 50, wherein the circulating microvesicles are exosomes or urinary exosomes.

52. The method of any one of embodiments 1-51, wherein the cell-free bodily fluid is separated from a bodily fluid.

53. The method of embodiment 52, wherein the bodily fluid is blood, urine, stool, saliva, lymph fluid, cerebrospinal fluid, synovial fluid, cystic fluid, ascites, pleural effusion, fluid obtained from a pregnant woman in the first trimester, fluid obtained from a pregnant woman in the second trimester, fluid obtained from a pregnant woman in the third trimester, maternal blood, amniotic fluid, chorionic villus sample, fluid from a preimplantation embryo, maternal urine, maternal saliva, placental sample, fetal blood, lavage and cervical vaginal fluid, interstitial fluid, buccal swab sample, sputum, bronchial lavage, Pap smear sample, or ocular fluid.

54. The method of any one of embodiments 52 and 53, wherein the cell-free bodily fluid is separated by filtration, centrifugation, flow cytometry, fluorescence activated cell sorting, gradient-based centrifugation, elution, microfluidics, magnetic separation technique, fluorescent-magnetic separation technique, nanostructure, quantum dots, high throughput microscope-based platforms, or a combination thereof.

55. The method of any one of embodiments 52-54, wherein the cell-free bodily fluid is separated by using a substance present in the sample.

56. The method of embodiment 55, wherein the substance is a product of a marker of said disease or condition.

57. The method of any one of embodiments 1-56, wherein the cell-free bodily fluid is plasma or serum.

58. The method of any one of embodiments 1-57, wherein the phagocytic cells, >2n phagocytic cells, =2n phagocytic cells, or non-phagocytic cells are isolated from a bodily fluid, tissues, or cells of the subject.

59. The method of embodiment 58, wherein the bodily fluid sample is blood, urine, stool, saliva, lymph fluid, cerebrospinal fluid, synovial fluid, cystic fluid, ascites, pleural effusion, fluid obtained from a pregnant woman in the first trimester, fluid obtained from a pregnant woman in the second trimester, fluid obtained from a pregnant woman in the third trimester, maternal blood, amniotic fluid, chorionic villus sample, fluid from a preimplantation embryo, maternal urine, maternal saliva, placental sample, fetal blood, lavage and cervical vaginal fluid, interstitial fluid, buccal swab sample, sputum, bronchial lavage, Pap smear sample, or ocular fluid.

60. The method of embodiment 58, wherein the cells are white blood cells.

61. The method of any one of embodiments 58-60, wherein the phagocytic cells, >2n phagocytic cells, =2n phagocytic cells, or non-phagocytic cells are isolated using antibodies.

62. The method of any one of embodiments 58-60, wherein the phagocytic cells, >2n phagocytic cells, =2n phagocytic cells, or non-phagocytic cells are isolated by flow cytometry, fluorescence activated cell sorting, filtration, gradient-based centrifugation, elution, microfluidics, magnetic separation technique, fluorescent-magnetic separation technique, nanostructure, quantum dots, high throughput microscope-based platforms, or a combination thereof.

63. The method of any one of embodiments 1-62, wherein the one or more markers are nucleic acids, proteins, lipids, carbohydrates, metabolites, or combinations thereof.

64. The method of embodiment 63, wherein the nucleic acids are nucleotides, oligonucleotides, DNAs, RNAs, or DNA-RNA hybrids.

65. The method of embodiment 64, wherein the DNAs are double-stranded DNAs, single-stranded DNAs, multi-stranded DNAs, complementary DNAs, genomic DNAs, or non-coding DNAs.

66. The method of embodiment 64, wherein the RNAs are messenger RNAs (mRNAs), microRNAs (miRNAs), small nucleolar RNAs (snoRNAs), ribosomal RNAs (rRNAs), transfer RNAs (tRNAs), small interfering RNAs (siRNAs), heterogeneous nuclear RNAs (hnRNAs), or small hairpin RNAs (shRNAs).

67. The method of embodiment 63, wherein the proteins are amino acids, peptides, enzymes, antigens, antibodies, cytokines, lipoproteins, glycoproteins, or hormones.

68. The method of embodiment 63, wherein the lipids are fatty acids, neutral fats, phosphatides, cholesterol, cholesterol esters, triglycerides, glycolipids, glycerolipids, glycerophospholipids, sphingolipids, sterol lipids, prenol lipids, saccharolipids, polyketides, choline glycerophospholipid, ethanolamine glycerophospholipid, phosphatidylinositol, phosphatidylglycerol, phosphatidylserine, lyso-choline glycerophospholipid, lyso-ethanolamine glycerophospholipid, phosphatidic acid, lyso-phosphatidic acid, sphingomyelin, galactosylceramide, glucosylceramide, free fatty acids, prostaglandins, triacylglycerol, diacylglycerol, monoacylglycerol, acyl-CoA, acylcarnitine, oxysterol, ceramide, cardiolipin, sphingoid base-1-phosphate, shingosine, lyso-sphingomyelin,, gangliosides, plasmalogen, sulfatide, low density lipoproteins (LDLs), very low density lipoproteins (VLDLs), high density lipoproteins (HDLs), sphingoid base-1-phosphates, or derivatives thereof.

69. The method of embodiment 63, wherein the carbohydrates are monosaccharides, disaccharides, polysaccharides, oligosaccharides, or derivatives thereof.

70. The method of embodiment 63, wherein the metabolites are primary metabolites, secondary metabolites, organic metabolites, inorganic metabolites, prostaglandins, hydroxyeicosatetraenoic acids, hydroxyoctadecadienoic acids, steroids, bile acids, vitamins, or derivatives thereof.

71. The method of any one of embodiments 1-3, 10-12, and 19-24, wherein the profile is a nucleic acid profile, a protein profile, a lipid profile, a carbohydrate profile, a metabolite profile, or a combination thereof.

72. The method of embodiment 65, wherein the profile is determined by a qualitative assay, a quantitative assay, or a combination thereof.

73. The method of any one of embodiments 4-9 and 13-18, wherein the first profile or the second profile is a nucleic acid profile, a protein profile, a lipid profile, a carbohydrate profile, a metabolite profile, or a combination thereof.

74. The method of embodiment 65, wherein the first profile or the second profile is determined by a qualitative assay, a quantitative assay, or a combination thereof.

75. The method of any one of embodiments 4-6 and 74, wherein the third profile or the fourth profile is a nucleic acid profile, a protein profile, a lipid profile, a carbohydrate profile, a metabolite profile, or a combination thereof.

76. The method of embodiment 75, wherein the third profile or the fourth profile is determined by a qualitative assay, a quantitative assay, or a combination thereof.

77. The method of any one of embodiments 72, 66, and 76, wherein the quantitative assay uses sequencing, targeted sequencing, single molecule real-time sequencing, electron microscopy-based sequencing, transistor-mediated sequencing, direct sequencing, random shotgun sequencing, Sanger dideoxy termination sequencing, exon sequencing, whole-genome sequencing, sequencing by hybridization, pyrosequencing, capillary electrophoresis, gel electrophoresis, duplex sequencing, cycle sequencing, single-base extension sequencing, solid-phase sequencing, high-throughput sequencing, massively parallel signature sequencing, emulsion PCR, multiplex PCR, co-amplification at lower denaturation temperature-PCR (COLD-PCR), sequencing by reversible dye terminator, paired-end sequencing, near-term sequencing, exonuclease sequencing, sequencing by ligation, short-read sequencing, single-molecule sequencing, sequencing-by-synthesis, real-time sequencing, reverse-terminator sequencing, nanopore sequencing, 454 sequencing, Solexa Genome Analyzer sequencing, SOLiD® sequencing, MS-PET sequencing,

mass spectrometry, matrix assisted laser desorption/ionization-time of flight (MALDI-TOF) mass spectrometry, electrospray ionization (ESI) mass spectrometry, surface-enhanced laser deorption/ionization-time of flight (SELDI-TOF) mass spectrometry, quadrupole-time of flight (Q-TOF) mass spectrometry, atmospheric pressure photoionization mass spectrometry (APPI-MS), Fourier transform mass spectrometry (FTMS), matrix-assisted laser desorption/ionization-Fourier transform-ion cyclotron resonance (MALDI-FT-ICR) mass spectrometry, secondary ion mass spectrometry (SIMS), polymerase chain reaction (PCR) analysis, co-amplification at lower denaturation temperature-PCR (COLD-PCR), multiplex PCR, quantitative PCR, real-time PCR, fluorescence assay, colorimetric assay, chemiluminescent assay, or a combination thereof.

78. The method of embodiment 77, wherein the nucleic acid profile is a genotypic profile, a single nucleotide polymorphism profile, a gene mutation profile, a gene copy number profile, a DNA methylation profile, a DNA acetylation profile, a chromosome dosage profile, a gene expression profile, or a combination thereof.

79. The method of embodiment 77, wherein the nucleic acid profile is determined by polymerase chain reaction (PCR) analysis, sequencing analysis, electrophoretic analysis, restriction fragment length polymorphism (RFLP) analysis, Northern blot analysis, quantitative PCR, reverse-transcriptase-PCR analysis (RT-PCR), allele-specific oligonucleotide hybridization analysis, comparative genomic hybridization, heteroduplex mobility assay (HMA), single strand conformational polymorphism (SSCP), denaturing gradient gel electrophisis (DGGE), RNAase mismatch analysis, mass spectrometry, tandem mass spectrometry, matrix assisted laser desorption/ionization-time of flight (MALDI-TOF) mass spectrometry, electrospray ionization (ESI) mass spectrometry, surface-enhanced laser deorption/ionization-time of flight (SELDI-TOF) mass spectrometry, quadrupole-time of flight (Q-TOF) mass spectrometry, atmospheric pressure photoionization mass spectrometry (APPI-MS), Fourier transform mass spectrometry (FTMS), matrix-assisted laser desorption/ionization-Fourier transform-ion cyclotron resonance (MALDI-FT-ICR) mass spectrometry, secondary ion mass spectrometry (SIMS), surface plasmon resonance, Southern blot analysis, in situ hybridization, fluorescence in situ

hybridization (FISH), chromogenic in situ hybridization (CISH), immunohistochemistry (IHC), microarray, comparative genomic hybridization, karyotyping, multiplex ligation-dependent probe amplification (MLPA), Quantitative Multiplex PCR of Short Fluorescent Fragments (QMPSF), microscopy, methylation specific PCR (MSP) assay, HpaII tiny fragment Enrichment by Ligation-mediated PCR (HELP) assay, radioactive acetate labeling assays, colorimetric DNA acetylation assay, chromatin immunoprecipitation combined with microarray (ChIP-on-chip) assay, restriction landmark genomic scanning, Methylated DNA immunoprecipitation (MeDIP), molecular break light assay for DNA adenine methyltransferase activity, chromatographic separation, methylation-sensitive restriction enzyme analysis, bisulfite-driven conversion of non-methylated cytosine to uracil, co-amplification at lower denaturation temperature-PCR (COLD-PCR), multiplex PCR, methyl-binding PCR analysis, or a combination thereof.

80. The method of embodiment 77, wherein the nucleic acid profile is determined by a sequencing technique selected from the group consisting of targeted sequencing, single molecule real-time sequencing, exon sequencing, electron microscopy-based sequencing, transistor-mediated sequencing, direct sequencing, random shotgun sequencing, Sanger dideoxy termination sequencing, whole-genome sequencing, sequencing by hybridization, pyrosequencing, capillary electrophoresis, gel electrophoresis, duplex sequencing, cycle sequencing, single-base extension sequencing, solid-phase sequencing, high-throughput sequencing, massively parallel signature sequencing, emulsion PCR, co-amplification at lower denaturation temperature-PCR (COLD-PCR), multiplex PCR, sequencing by reversible dye terminator, paired-end sequencing, near-term sequencing, exonuclease sequencing, sequencing by ligation, short-read sequencing, single-molecule sequencing, sequencing-by-synthesis, real-time sequencing, reverse-terminator sequencing, nanopore sequencing, 454 sequencing, Solexa Genome Analyzer sequencing, SOLiD® sequencing, MS-PET sequencing, mass spectrometry, and a combination thereof.

81. The method of embodiment 77, wherein the protein profile is a protein expression profile, a protein activation profile, or a combination thereof.

82. The method of embodiment 77, wherein the protein profile is determined by an immunohistochemistry assay, an enzyme-linked immunosorbent assay (ELISA), in situ hybridization, chromatography, liquid chromatography, size exclusion chromatography, high performance liquid chromatography (HPLC), gas chromatography, mass spectrometry, tandem mass spectrometry, matrix assisted laser desorption/ionization-time of flight (MALDI-TOF) mass spectrometry, electrospray ionization (ESI) mass spectrometry, surface-enhanced laser deorption/ionization-time of flight (SELDI-TOF) mass spectrometry, quadrupole-time of flight (Q-TOF) mass spectrometry, atmospheric pressure photoionization mass spectrometry (APPI-MS), Fourier transform mass spectrometry (FTMS), matrix-assisted laser desorption/ionization-Fourier transform-ion cyclotron resonance (MALDI-FT-ICR) mass spectrometry, secondary ion mass spectrometry (SIMS), radioimmunoassays, microscopy, microfluidic chip-based assays, surface plasmon resonance, sequencing, Western blotting assay, or a combination thereof.

83. The method of embodiment 77, wherein the protein activation profile comprises determining a phosphorylation state, an ubiquitination state, a myristoylation state, a conformational state, or a combination thereof of the one or more markers.

84. The method of embodiment 77, wherein the lipid profile is determined by chromatography, liquid chromatography, size exclusion chromatography, high performance liquid chromatography (HPLC), gas chromatography, mass spectrometry, tandem mass spectrometry, matrix assisted laser desorption/ionization-time of flight (MALDI-TOF) mass spectrometry, electrospray ionization (ESI) mass spectrometry, surface-enhanced laser deorption/ionization-time of flight (SELDI-TOF) mass spectrometry, quadrupole-time of flight (Q-TOF) mass spectrometry, atmospheric pressure photoionization mass spectrometry (APPI-MS), Fourier transform mass spectrometry (FTMS), matrix-assisted laser desorption/ionization-Fourier transform-ion cyclotron resonance (MALDI-FT-ICR) mass spectrometry, secondary ion mass spectrometry (SIMS), radioimmunoassays, microfluidic chip-based assay, detection of fluorescence, detection of chemiluminescence, or a combination thereof.

85. The method of embodiment 77, wherein the carbohydrate profile is determined by chromatography, liquid chromatography, size exclusion chromatography, high performance anion exchange chromatography with pulsed amperometric detection (HPAEC-PAD), liquid chromatography, gas chromatography, fluorescent assay, mass spectrometry, tandem mass spectrometry, matrix assisted laser desorption/ionization-time of flight (MALDI-TOF) mass spectrometry, electrospray ionization (ESI) mass spectrometry, surface-enhanced laser deorption/ionization-time of flight (SELDI-TOF) mass spectrometry, quadrupole-time of flight (Q-TOF) mass spectrometry, atmospheric pressure photoionization mass spectrometry (APPI-MS), Fourier transform mass spectrometry (FTMS), matrix-assisted laser desorption/ionization-Fourier transform-ion cyclotron resonance (MALDI-FT-ICR) mass spectrometry, secondary ion mass spectrometry (SIMS), radioimmunoassay, microfluidic chip-based assay, detection of fluorescence, detection of chemiluminescence, or a combination thereof.

86. The method of any one of embodiments 1-85, wherein the subject has at least two diseases or conditions.

87. The method of embodiment 86, wherein the subject has at least one prenatal or pregnancy-related disease or condition

88. The method of any one of embodiments 1-87, wherein the subject is a mammal.

89. The method of embodiment 89, wherein the mammal is a human.

90. The method of any one of embodiments 1-89, wherein the disease or condition is a cardiovascular disease or condition, a kidney-associated disease or condition, a prenatal or pregnancy-related disease or condition, a neurological or neuropsychiatric disease or condition, an autoimmune or immune-related disease or condition, a cancer, an infectious disease or condition, a mitochondrial disorder, a respiratory-gastrointestinal tract disease or condition, a reproductive disease or condition, an ophthalmic disease or condition, a musculo-skeletal disease or condition, or a dermal disease or condition.

91. The method of any one of embodiments 1-90, wherein the difference is greater than a 1-fold difference.

92. The method of embodiment 91, wherein the difference is at least 1.05-fold, 1.1-fold, 1.2-fold, 1.3-fold, 1.4-fold, 1.5-fold, 2-fold, 2.5-fold, 3-fold, 4-fold, 5-fold, 6-fold, 7-fold, 8-fold, 9-fold, or 10-fold difference.

93. The method of any one of embodiments 4-6 and 13-15, wherein the second profile and the fourth profile are the same.

94. A method for identifying one or more markers for a disease or condition comprising:

a) determining a first profile of analytes from a sample comprising a cell-free bodily fluid from a subject having said disease or condition and a population of phagocytic cells, or a population of >2n phagocytic cells, from a subject having said disease or condition;

determining a second profile of analytes from a population of =2n phagocytic cells, or a population of non-phagocytic cells, from the subject having said disease or condition;

identifying a set of differences between the first and second profiles, wherein the first set of differences is specific to the first profile relative to the second profile;

b) determining a third profile of analytes from a sample comprising a cell-free bodily fluid from a subject having said disease or condition and a population of phagocytic cells, or a population of >2n phagocytic cells, from a control subject not having said disease or condition;

determining a fourth profile of analytes from a population of =2n phagocytic cells, or a population of non-phagocytic cells, from the control subject not having said disease or condition;

identifying a set of differences between the third and fourth profiles, wherein the second set of differences is specific to the third profile relative to the fourth profile; and

c) identifying one or more analytes specific to the set of differences identified in a) relative to the set of differences identified in b), the identified analytes in c) being markers of said disease or condition.

95. A method for identifying one or more markers for a disease or condition comprising:

a) determining a first profile of analytes from a sample comprising a cell-free bodily fluid from a subject having said disease or condition and a population of phagocytic cells, or a population of >2n phagocytic cells, from a subject having said disease or condition;

b) comparing the first profile to a second profile derived from a repository of analytes from a control subject not having said disease or condition;

c) identifying a set of differences between the first and second profiles, wherein the set of differences is specific to the first profile relative to the second profile; and

d) identifying one or more analytes specific to the set of differences, the identified analytes being markers of said disease or condition.

96. The method of embodiment 95, further comprising:

a) obtaining a fifth profile of analytes from cells or tissues affected by said disease or condition in the subject having said disease or condition;

obtaining a sixth profile of analytes from cells or tissues not affected by said disease or condition in the subject having said disease or condition;

identifying a set of differences between the fifth and sixth profiles, wherein the set of differences is specific to the fifth profile relative to the sixth profile; and

b) identifying at least one of the one or more markers of c) present in the set of differences identified in d).

97. The method of any one of embodiments 90-96, further comprising extracting the one or more markers from the cell-free bodily fluid.

98. The method of any one of embodiments 90-97, further comprising lysing the phagocytic cells, the >2n phagocytic cells, the =2n phagocytic cells, or the non-phagocytic cells before a).

99. The method of any one of embodiments 90-98, further comprising extracting at least some of the cellular contents from the phagocytic cells, the >2n phagocytic cells, the =2n phagocytic cells, or the non-phagocytic cells before a).

100. The method of any one of embodiments 90-99, wherein the phagocytic cells or the >2n phagocytic cells comprise viable diseased cells, dead diseased cells, apoptotic diseased cells, circulating tumor cells, infectious agents, fetal cells, trophoblasts, or fragments thereof.

101. The method of any one of embodiments 90-100, wherein at least one of the one or more markers of said disease or condition is present in the cell-free bodily fluid sample, the phagocytic cells, or the >2n phagocytic cells.

102. The method of any one of embodiments 90-101, wherein at least one of the one or more markers of said disease or condition is not present in the cell-free bodily fluid sample, the phagocytic cells, or the >2n phagocytic cells.

103. The method of any one of embodiments 90-102, further comprising comparing the identified difference of c) to a repository of one or more known markers of said disease or condition.

104. The method of embodiment 97, wherein the repository is obtained by data mining.

105. The method of any one of embodiments 90-104, wherein the phagocytic cells, >2n phagocytic cells, or =2n phagocytic cells are neutrophils, macrophages, monocytes, dendritic cells, foam cells, mast cells, eosinophils, keratinocytes, or mixtures thereof.

106. The method of any one of the embodiments 90-105, wherein the non-phagocytic cells are T cells, B cells, null cells, basophils, or mixtures thereof

107. The method of any one of embodiments 90-106, wherein the cell-free bodily fluid is separated from a bodily fluid.

108. The method of embodiment 107, wherein the bodily fluid sample is blood, urine, stool, saliva, lymph fluid, cerebrospinal fluid, synovial fluid, cystic fluid, ascites, pleural effusion, fluid obtained from a pregnant woman in the first trimester, fluid obtained from a pregnant woman in the second trimester, fluid obtained from a pregnant woman in the third trimester, maternal blood, amniotic fluid, chorionic villus sample, fluid from a preimplantation embryo, maternal urine, maternal saliva, placental sample, fetal blood, lavage and cervical vaginal fluid, interstitial fluid, or ocular fluid.

109. The method of any one of embodiments 107 and 108, wherein the cell-free bodily fluid sample is separated by filtration, centrifugation, flow cytometry, fluorescence activated cell sorting, gradient-based centrifugation, elution, microfluidics, magnetic separation technique, fluorescent-magnetic separation technique, nanostructure, quantum dots, high throughput microscope-based platforms, or a combination thereof.

110. The method of embodiment 109, wherein the cell-free bodily fluid sample is separated by using a substance present in the sample.

111. The method of any one of embodiments 90-110, wherein the one or more markers are nucleic acids, proteins, lipids, carbohydrates, metabolites, or combinations thereof.

112. The method of embodiment 111, wherein the nucleic acids are nucleotides, oligonucleotides, DNAs, RNAs, or DNA-RNA hybrids.

113. The method of embodiment 112, wherein the DNAs are double-stranded DNAs, single-stranded DNAs, multi-stranded DNAs, complementary DNAs, genomic DNAs, or non-coding DNAs.

114. The method of embodiment 112, wherein the RNAs are messenger RNAs (mRNAs), microRNAs (miRNAs), small nucleolar RNAs (snoRNAs), ribosomal RNAs (rRNAs), transfer RNAs (tRNAs), small interfering RNAs (siRNAs), heterogeneous nuclear RNAs (hnRNAs), or small hairpin RNAs (shRNAs).

115. The method of embodiment 111, wherein the proteins are amino acids, peptides, enzymes, antigens, antibodies, cytokines, lipoproteins, glycoproteins, or hormones.

116. The method of embodiment 111, wherein the lipids are fatty acids, neutral fats, phosphatides, cholesterol, cholesterol esters, triglycerides, glycolipids, glycerolipids, glycerophospholipids, sphingolipids, sterol lipids, prenol lipids, saccharolipids, polyketides, choline glycerophospholipid, ethanolamine glycerophospholipid, phosphatidylinositol, phosphatidylglycerol, phosphatidylserine, lyso-choline glycerophospholipid, lyso-ethanolamine glycerophospholipid, phosphatidic acid, lyso-phosphatidic acid, sphingomyelin, galactosylceramide, glucosylceramide, free fatty acids, prostaglandins, triacylglycerol, diacylglycerol, monoacylglycerol, acyl-CoA, acylcarnitine, oxysterol, ceramide, cardiolipin, sphingoid base-1-phosphate, shingosine, lyso-sphingomyelin,, gangliosides, plasmalogen, sulfatide, low density lipoproteins (LDLs), very low density lipoproteins (VLDLs), high density lipoproteins (HDLs), sphingoid base-1-phosphates, or derivatives thereof.

117. The method of embodiment 111, wherein the carbohydrates are monosaccharides, disaccharides, polysaccharides, oligosaccharides, or derivatives thereof.

118. The method of embodiment 111, wherein the metabolites are primary metabolites, secondary metabolites, organic metabolites, inorganic metabolites, prostaglandins, hydroxyeicosatetraenoic acids, hydroxyoctadecadienoic acids, steroids, bile acids, vitamins, or derivatives thereof.

119. The method of any one of embodiments 90-118, wherein the profile is a nucleic acid profile, a protein profile, a lipid profile, a carbohydrate profile, a metabolite profile, or a combination thereof.

120. The method of embodiment 119, wherein the profile is determined by a qualitative assay, a quantitative assay, or a combination thereof.

121. The method of embodiment 120, wherein the quantitative assay uses sequencing, targeted sequencing, single molecule real-time sequencing, exon sequencing, electron microscopy-based sequencing, transistor-mediated sequencing, direct sequencing, random shotgun sequencing, Sanger dideoxy termination sequencing, whole-genome sequencing, sequencing by hybridization, pyrosequencing, capillary electrophoresis, gel electrophoresis, duplex sequencing, cycle sequencing, single-base extension sequencing, solid-phase sequencing, high-throughput sequencing, massively parallel signature sequencing, emulsion PCR, co-amplification at lower denaturation temperature-PCR (COLD-PCR), multiplex PCR, sequencing by reversible dye terminator, paired-end sequencing, near-term sequencing, exonuclease sequencing, sequencing by ligation, short-read sequencing, single-molecule sequencing, sequencing-by-synthesis, real-time sequencing, reverse-terminator sequencing, nanopore sequencing, 454 sequencing, Solexa Genome Analyzer sequencing, SOLiD® sequencing, MS-PET sequencing, mass spectrometry, matrix assisted laser desorption/ionization-time of flight (MALDI-TOF) mass spectrometry, electrospray ionization (ESI) mass spectrometry, surface-enhanced laser desorption/ionization-time of flight (SELDI-TOF) mass spectrometry, quadrupole-time of flight (Q-TOF) mass spectrometry, atmospheric pressure photoionization mass

spectrometry (APPI-MS), Fourier transform mass spectrometry (FTMS), matrix-assisted laser desorption/ionization-Fourier transform-ion cyclotron resonance (MALDI-FT-ICR) mass spectrometry, secondary ion mass spectrometry (SIMS), polymerase chain reaction (PCR) analysis, quantitative PCR, real-time PCR, fluorescence assay, colorimetric assay, chemiluminescent assay, or a combination thereof.

122. The method of embodiment 119, wherein the nucleic acid profile is a genotypic profile, a single nucleotide polymorphism profile, a gene mutation profile, a gene copy number profile, a DNA methylation profile, a DNA acetylation profile, a chromosome dosage profile, a gene expression profile, or a combination thereof.

123. The method of embodiment 119, wherein the nucleic acid profile is determined by polymerase chain reaction (PCR) analysis, sequencing analysis, electrophoretic analysis, restriction fragment length polymorphism (RFLP) analysis, Northern blot analysis, quantitative PCR, reverse-transcriptase-PCR analysis (RT-PCR), allele-specific oligonucleotide hybridization analysis, comparative genomic hybridization, heteroduplex mobility assay (HMA), single strand conformational polymorphism (SSCP), denaturing gradient gel electrophisis (DGGE), RNAase mismatch analysis, mass spectrometry, tandem mass spectrometry, matrix assisted laser desorption/ionization-time of flight (MALDI-TOF) mass spectrometry, electrospray ionization (ESI) mass spectrometry, surface-enhanced laser deorption/ionization-time of flight (SELDI-TOF) mass spectrometry, quadrupole-time of flight (Q-TOF) mass spectrometry, atmospheric pressure photoionization mass spectrometry (APPI-MS), Fourier transform mass spectrometry (FTMS), matrix-assisted laser desorption/ionization-Fourier transform-ion cyclotron resonance (MALDI-FT-ICR) mass spectrometry, secondary ion mass spectrometry (SIMS), surface plasmon resonance, Southern blot analysis, in situ hybridization, fluorescence in situ hybridization (FISH), chromogenic in situ hybridization (CISH), immunohistochemistry (IHC), microarray, comparative genomic hybridization, karyotyping, multiplex ligation-dependent probe amplification (MLPA), Quantitative Multiplex PCR of Short Fluorescent Fragments (QMPSF), microscopy, methylation specific PCR (MSP) assay, HpaII tiny fragment Enrichment by Ligation-mediated

PCR (HELP) assay, radioactive acetate labeling assays, colorimetric DNA acetylation assay, chromatin immunoprecipitation combined with microarray (ChIP-on-chip) assay, restriction landmark genomic scanning, Methylated DNA immunoprecipitation (MeDIP), molecular break light assay for DNA adenine methyltransferase activity, chromatographic separation, methylation-sensitive restriction enzyme analysis, bisulfite-driven conversion of non-methylated cytosine to uracil, co-amplification at lower denaturation temperature-PCR (COLD-PCR), multiplex PCR, methyl-binding PCR analysis, or a combination thereof.

124. The method of embodiment 119, wherein the nucleic acid profile is determined by a sequencing technique selected from the group consisting of targeted sequencing, single molecule real-time sequencing, exon sequencing, electron microscopy-based sequencing, transistor-mediated sequencing, direct sequencing, random shotgun sequencing, Sanger dideoxy termination sequencing, whole-genome sequencing, sequencing by hybridization, pyrosequencing, capillary electrophoresis, gel electrophoresis, duplex sequencing, cycle sequencing, single-base extension sequencing, solid-phase sequencing, high-throughput sequencing, massively parallel signature sequencing, emulsion PCR, co-amplification at lower denaturation temperature-PCR (COLD-PCR), multiplex PCR, sequencing by reversible dye terminator, paired-end sequencing, near-term sequencing, exonuclease sequencing, sequencing by ligation, short-read sequencing, single-molecule sequencing, sequencing-by-synthesis, real-time sequencing, reverse-terminator sequencing, nanopore sequencing, 454 sequencing, Solexa Genome Analyzer sequencing, SOLiD® sequencing, MS-PET sequencing, mass spectrometry, and a combination thereof.

125. The method of embodiment 119, wherein the protein profile is a protein expression profile, a protein activation profile, or a combination thereof.

126. The method of embodiment 119, wherein the protein profile is determined by an immunohistochemistry assay, an enzyme-linked immunosorbent assay (ELISA), in situ hybridization, chromatography, liquid chromatography, size exclusion chromatography, high performance liquid chromatography (HPLC), gas

chromatography, mass spectrometry, tandem mass spectrometry, matrix assisted laser desorption/ionization-time of flight (MALDI-TOF) mass spectrometry, electrospray ionization (ESI) mass spectrometry, surface-enhanced laser deorption/ionization-time of flight (SELDI-TOF) mass spectrometry, quadrupole-time of flight (Q-TOF) mass spectrometry, atmospheric pressure photoionization mass spectrometry (APPI-MS), Fourier transform mass spectrometry (FTMS), matrix-assisted laser desorption/ionization-Fourier transform-ion cyclotron resonance (MALDI-FT-ICR) mass spectrometry, secondary ion mass spectrometry (SIMS), radioimmunoassays, microscopy, microfluidic chip-based assays, surface plasmon resonance, sequencing, Western blotting assay, or a combination thereof.

127. The method of embodiment 119, wherein the protein activation profile comprises determining a phosphorylation state, an ubiquitination state, a myristoylation state, a conformational state, or a combination thereof of the one or more markers.

128. The method of embodiment 119, wherein the lipid profile is determined by chromatography, liquid chromatography, size exclusion chromatography, high performance liquid chromatography (HPLC), gas chromatography, mass spectrometry, tandem mass spectrometry, matrix assisted laser desorption/ionization-time of flight (MALDI-TOF) mass spectrometry, electrospray ionization (ESI) mass spectrometry, surface-enhanced laser deorption/ionization-time of flight (SELDI-TOF) mass spectrometry, quadrupole-time of flight (Q-TOF) mass spectrometry, atmospheric pressure photoionization mass spectrometry (APPI-MS), Fourier transform mass spectrometry (FTMS), matrix-assisted laser desorption/ionization-Fourier transform-ion cyclotron resonance (MALDI-FT-ICR) mass spectrometry, secondary ion mass spectrometry (SIMS), radioimmunoassays, microfluidic chip-based assay, detection of fluorescence, detection of chemiluminescence, or a combination thereof.

129. The method of embodiment 119, wherein the carbohydrate profile is determined by chromatography, liquid chromatography, size exclusion chromatography, high performance anion exchange chromatography with pulsed

amperometric detection (HPAEC-PAD), liquid chromatography, gas chromatography, fluorescent assay, mass spectrometry, tandem mass spectrometry, matrix assisted laser desorption/ionization-time of flight (MALDI-TOF) mass spectrometry, electrospray ionization (ESI) mass spectrometry, surface-enhanced laser deorption/ionization-time of flight (SELDI-TOF) mass spectrometry, quadrupole-time of flight (Q-TOF) mass spectrometry, atmospheric pressure photoionization mass spectrometry (APPI-MS), Fourier transform mass spectrometry (FTMS), matrix-assisted laser desorption/ionization-Fourier transform-ion cyclotron resonance (MALDI-FT-ICR) mass spectrometry, secondary ion mass spectrometry (SIMS), radioimmunoassay, microfluidic chip-based assay, detection of fluorescence, detection of chemiluminescence, or a combination thereof.

130. The method of any one of embodiments 90-129, wherein the subject is a mammal

131. The method of embodiment 130, wherein the mammal is a human.

132. The method of any one of embodiments 90-131, wherein the disease or condition is a cardiovascular disease or condition, a kidney-associated disease or condition, a prenatal or pregnancy-related disease or condition, a neurological or neuropsychiatric disease or condition, an autoimmune or immune-related disease or condition, a cancer, an infectious disease or condition, a mitochondrial disorder, a respiratory-gastrointestinal tract disease or condition, a reproductive disease or condition, an ophthalmic disease or condition, a musculo-skeletal disease or condition, or a dermal disease or condition.

133. The method of any one of embodiments 90-132, wherein the difference is greater than a 1-fold difference.

134. The method of embodiment 133, wherein the difference is at least 1.05-fold, 1.1-fold, 1.2-fold, 1.3-fold, 1.4-fold, 1.5-fold, 2-fold, 2.5-fold, 3-fold, 4-fold, 5-fold, 6-fold, 7-fold, 8-fold, 9-fold, or 10-fold difference.

135. The method of any one of embodiments 1-93, further comprising determining at least one diagnostic parameter of said disease or condition.

136. The method of embodiment 135, wherein the diagnostic parameter is determined by physical inspection, visual inspection, biopsy, scanning, histology, radiology, imaging, ultrasound, use of a commercial kit, genetic testing, immunological testing, analysis of bodily fluids, or monitoring neural activity.

137. The method of any one of the embodiments 1-93 and 135-136, wherein the one or more markers comprise at least one gene selected from the group consisting of AKT2, BAK1, EGFR, ERBB2, ETS2, FOS, JUN, MAP2K1, MMP2, PDGFB, RB1, SERPINB2, SNCG, and SPP1.

138. The method of any one of the embodiments 1-93 and 135-137, wherein the one or more markers comprise at least one gene selected from the group consisting of AKT1, AKT2, BAK2, CDC25A, E2F1, EGFR, ERBB2, FOS, JUN, MAP2K1, MMP2, NFKB1, PDGFB, PIK3R1, PNN, RB1, SERPINB2, SERPINB5, SNCG, SPP1, TERT, TIMP3, and TP53.

139. The method of any one of the embodiments 1-93 and 135-138, wherein the one or more markers comprise at least one gene selected from the group consisting of CASP8, CASP9, COL18A1, ETS2, HTATIP2, MMP9, SRC, and TWIST1.

140. The method of any one of the embodiments 1-93 and 135-139, wherein the one or more markers comprise at least one gene selected from the group consisting of AKT1, APAF1, ATM, CDC25A, CDKN1A, ETS2, FOS, IL8, ITGA4, ITGA6, ITGAV, JUN, MAP2K1, NFKBIA, PLAU, PLAUR, RAF1, SERPINB2, SYK, TIMP1, TNF, TNFRSF10B, and TNFRSF1A.

141. The method of any one of the embodiments 1-93 and 135-140, wherein the one or more markers comprise at least one gene selected from the group consisting of ACP2, AK2, AKT3, ARL5B, ATP2B3, BGN, BRAF, BTG2, CAMKK2, CAPG, CAPN12, CPLX2, DENND5A, DNA2, FAM104A, FNIP1, GFRA4, GLUD1, GNAQ, GP1BB, HNRPLL, HOXA2, HPS3, INPP4A, ITGAV, KLHL23, LANCL2, LYPD6, MAPKAPK3, MEF2A (includes, EG:4205), MEF2C, NVL, PCYT1A, PGLYRP4, PLOD1, PPP1CB, PRKAB2, PROS1, PTPRE, RASA4

(includes,EG:10156), RBMS2, RBPJ, STAT5B, THBS1, TRIB1, TRIM2, TSPAN6, and ZDHHC21.

142. The method of any one of the embodiments 1-93 and 135-141, wherein the one or more markers comprise at least one gene selected from the group consisting of B4GALT5, BOP1, CCL2, CCL3, CCL3L1, CCRL2, CD83, CLEC4G, CLIC4, CTSC, CTSO, CXCL10, FCGR3A, FPR3, HBA1, HBB, LRMP, MAP1LC3B2, MS4A4A, MSR1, MYADML, NID1, PF4, PION, RNF217, SAMD9L, SERPING1, and SPARC.

143. The method of any one of the embodiments 1-93 and 135-142, wherein the one or more markers comprise at least one gene selected from the group consisting of ACOT9, AMPD2, ARHGAP15, BATF2, C3AR1, C5orf41, CCL3, CCL3L1, CD63, CHST11, CHSY1, CLEC4G, CTSZ, CXorf21, CYTH4, CYTIP, DLEU2, DNAJA1, DOCK8, DTX3L, DUSP6, EPSTI1, ERF, F2RL1, FYB, GABRB2, GBP5, GLRX, GNB4, ICAM1, IFI35, IFIH1, IFNAR2, IL1R1, IRF1, ITGA5, LAP3, LAPTM5, LCP2, MAP1LC3B, MAP1LC3B2, MICAL2, MT1DP, MT1JP, MT1M, MT2A, MYADML, NEK6, NINJ2, NNMT, NT5C3L, NUB1, PDE4B, PLOD1, PML, PRKCB, PSMB9, RCN3, RGS4, RNASE6, RTP4, SAMD9L, SEL1L, SERPING1, SETX, SIGLEC10, SKIL, SLC7A7, SNORA21, SP100, SP110, SP140, SSFA2, STAT2, STK17B, STK3, TDRD7, TMCC1, TMPRSS11E2, TNFRSF1B, TPM1, TRIM21,TXNDC4, UBE2L6, UBE2W, USP18, VAV1, WARS, WIPF1, and WIPI1.

144. The method of any one of the embodiments 1-93 and 135-143, wherein the one or more markers comprise at least one gene selected from the group consisting of ADAR, ADM, ALAS1, ANKRD22, ARHGAP27, B3GNT5, BCL10, C12orf35, C15orf29, C2orf59, CD177, CEACAM1, CPEB2, DDX58, F2RL1, GDPD3, GNAI3, HIST2H3A, HIST2H3D, HIST2H4A, HMGCR, HSPA6, HSPC159, IL4R, IMPA2, KPNB1, KREMEN1, KRT23, LDLR, LOC100130904, LTB4R, MAEA, MARK2, MBOAT2, MPZL3, N4BP1, NBEAL2, NMI, NPEPPS, PARP14, PGM2, PPIF, PXN, RALBP1, ROD1, RPS6KA1, S100P, SERTAD2, SLC9A1, SLPI, SP110, SPINT1, ST14, TBC1D3, TNFRSF9, TRIM21, UPP1, VPS24, ZBTB34, and ZNF256.

145. The method of any one of the embodiments 1-93 and 135-144, wherein the one or more markers comprise at least one or more of the markers identified by the methods of any one the embodiments 94-134.

146. A kit comprising a plurality of marker detection agents that detect at least one or more of the markers identified by the methods of any one of the embodiments 94-134.

147. A method of treating or preventing a disease or condition in a subject comprising administering to said subject a composition comprising a compound identified by the method of any one of embodiments 6 and 15.

148. The method of any one of the embodiments 1-93 and 135-145, wherein the circulating diseased cells are infected by an infectious agent.

149. The method of embodiment 148, wherein the infectious agent is a virus, bacteria, fungus, parasite, protozoan, infectious protein or microorganism.

150. The method of any one of embodiments 1-93, 135-145, and 148-149, wherein, when said method comprises phagocytic cells or >2n phagocytic cells, said phagocytic cells or >2n phagocytic cells comprise a transrenal nucleic acid.

Brief Description of the Drawings

[0008] Figure 1 depicts a set of potentially informative markers and the relative percentage of reads of associated allelic frequencies in maternal reference samples of CD2-positive peripheral blood cells and combination samples. Along the x-axis are the markers and along the y-axis are the percent reads for the different alleles. Triangle: percent of maternal sample reference allele; circle: percent of maternal sample alternate allele; square: percent of combination sample reference allele; and diamond: percent of combination sample alternate allele.

[0009] Figure 2 depicts a set of potentially informative markers and the relative percentage of reads of associated allelic frequencies in maternal reference samples and combination samples. Along the x-axis are the markers and along the y-axis are the percent reads for the different alleles. Triangle: percent of maternal sample

reference allele; circle: percent of maternal sample alternate allele; square: percent of combination sample reference allele; and diamond: percent of combination sample alternate allele.

[0010] Figure 3 depicts a set of potentially informative markers and the relative percentage of reads of associated allelic frequencies as compared between a circulating cell-free fluid (ccff) sample and a maternal reference sample. Along the x-axis are the markers and along the y-axis are the percent reads for the different alleles. Triangle: reference allele; circle: alternative allele.

[0011] Figure 4 depicts a set of potentially informative markers and the relative percentage of reads of associated allelic frequencies in a T-cell sample. Along the x-axis are the markers and along the y-axis are the percent reads for the different alleles. Triangle: reference allele; circle: alternative allele.

[0012] Figure 5 depicts a set of potentially informative markers and the relative percentage of reads of associated allelic frequencies in a combination sample of a circulating cell-free fluid (ccff) plus 1% monocyte sample. Along the x-axis are the markers and along the y-axis are the percent reads for the different alleles. Triangle: reference allele; circle: alternative allele.

[0013] Figure 6 depicts the emergence of informative markers in a circulating cell-free fluid fraction of the maternal blood compared to the lack of informative markers in the maternal T-cells. Along the x-axis are the markers and along the y-axis are the percent reads for the different alleles.

[0014] Figure 7 depicts the emergence of the informative markers in a combination sample containing cell-free bodily fluid plus monocytes from maternal blood compared to a lack of informative markers in the maternal reference sample containing T-cells. Along the x-axis are the markers and along the y-axis are the percent reads for the different alleles.

[0015] Figure 8 depicts an exemplary informative marker in a maternal only sample and a combination sample. Triangle: reference allele; circle: alternative allele.

[0016] Figure 9 depicts an exemplary informative marker in a maternal only sample and a combination sample. Triangle: reference allele; circle: alternative allele.

[0017] Figure 10 depicts an exemplary non-informative marker in a maternal only sample and a combination sample. Triangle: reference allele; circle: alternative allele.

Detailed Description of the Invention

[0018] Unless otherwise defined herein, scientific and technical terms used in this application shall have the meanings that are commonly understood by those of ordinary skill in the art. Generally, nomenclature used in connection with, and techniques of, cell and tissue culture, molecular biology, cell and cancer biology, neurobiology, neurochemistry, virology, immunology, microbiology, pharmacology, genetics and protein and nucleic acid chemistry, described herein, are those well known and commonly used in the art.

[0019] All of the above, and any other publications, patents and published patent applications referred to in this application are specifically incorporated by reference herein. In case of conflict, the present specification, including its specific definitions, will control.

[0020] Throughout this specification, the word "comprise" or variations such as "comprises" or "comprising" will be understood to imply the inclusion of a stated integer (or components) or group of integers (or components), but not the exclusion of any other integer (or components) or group of integers (or components).

[0021] The singular forms "a," "an," and "the" include the plurals unless the context clearly dictates otherwise.

[0022] The term "including" is used to mean "including but not limited to". "Including" and "including but not limited to" are used interchangeably.

[0023] A "patient", "subject", or "individual" are used interchangeably and refer to either a human or a non-human animal. These terms include mammals, such as humans, primates, livestock animals (e.g., bovines, porcines), companion animals (e.g., canines, felines) and rodents (e.g., mice and rats).

[0024] As used herein, a control subject refers to any individual that has not been diagnosed as having the disease or condition being assayed. The terms "normal control", "healthy control", and "not-diseased cells" likewise mean a sample (e.g., cells, serum, tissue) taken from a source (e.g., subject, control subject, cell line) that does not have the condition or disease being assayed and therefore may be used to determine the baseline for the condition or disorder being measured. It is also understood that the control subject, normal control, and healthy control, include data obtained and used as a standard, i.e. it can be used over and over again for multiple different subjects. In other words, for example, when comparing a subject sample to a control sample, the data from the control sample could have been obtained in a different set of experiments, for example, it could be an average obtained from a number of healthy subjects and not actually obtained at the time the data for the subject was obtained.

[0025] The term "diagnosis" as used herein refers to methods by which the skilled artisan can estimate and/or determine whether or not a patient is suffering from a given disease or condition. The skilled artisan often makes a diagnosis on the basis of one or more diagnostic indicators, e.g., a marker, the presence, absence, amount, or change in amount of which is indicative of the presence, severity, or absence of the condition.

[0026] The term "prognosis" as used herein refers to is used herein to refer to the likelihood of a disease or condition progression, including recurrence of a disease or condition.

[0027] The disclosure of the International Applications PCT/US09/31395, PCT/US11/45009, PCT/US11/44969, PCT/US11/44973, PCT/US11/44991, PCT/US11/45002, PCT/US11/44996, and PCT/US11/45018 are incorporated herein by reference for all purposes.

[0028] Each embodiment described herein may be combined with any other embodiment described herein.

Description of Methods of the Invention

[0029] By using a profile derived from a combination of sources, the skilled worker is able to capture data that is normally lost in the process of separating components out of a sample for testing. At the same time, the analytical components being combined are enriched for the markers used. Thus, the methods of the invention do not introduce unnecessary "noise" into the signal. Embodiments using a subject-specific profile comparison also eliminate the dependence on a population-derived average profile for a particular disease or condition. Using a population-derived average profile may introduce error into the detection or diagnosis of a particular disease or condition in the subject. Therefore, methods of this invention allow detection, diagnosis, and treatment to be personalized to the individual.

[0030] The methods of this invention have high specificity, sensitivity, and accuracy, and are capable of detecting disease or condition-specific markers present within a bodily fluid sample, cells or tissues. In some embodiments, the methods of this invention have improved specificity, sensitivity, and accuracy compared to current methods. The methods of the invention also reduce the problem of losing valuable signal when, for example, a patient sample is collected and separated into individual components, of which only one is chosen for testing. Accordingly, in certain aspects, the invention provides non-invasive assays for the early detection of a disease or condition, i.e., before the disease can be diagnosed by conventional diagnostic techniques, e.g., imaging techniques, and, therefore, provide a foundation for improved decision-making relative to the needs and strategies for intervention, prevention, and treatment of individuals with such disease or condition.

[0031] The present invention provides methods for diagnosing or aiding in the diagnosis of a disease or condition by comparing profiles (e.g., gene/protein/lipid/carbohydrate expression profiles, genotypes, gene copy number, gene dosage, DNA methylation, etc.) of disease or condition-associated markers (e.g., nucleic acids, proteins, lipids, carbohydrates, metabolites). One profile used for the comparison may be a profile from a sample comprising combinations of two or more different components (e.g., cell-free bodily fluids, phagocytic cells, circulating vesicles, and circulating diseased cells) from a subject. Alternatively, profile used for the comparison may be a profile from a sample comprising analytes isolated from,

e.g., cell-free bodily fluids, phagocytic cells, phagocytic cells having a DNA content more than 2n (>2n phagocytic cells), circulating vesicles, and circulating diseased cells from a subject. For example, a sample comprising analytes from phagocytes and a cell-free bodily fluid could be generated by isolating analytes (e.g., nucleic acids or proteins) from phagocytes (e.g., by lysing the cells and using affinity-based techniques to isolate the analytes) and analytes (e.g., nucleic acids or proteins) from a cell-free bodily fluid, and combining the analytes to create a sample useful in creating a profile. For ease of reference, a sample comprising combinations of two or more different components (e.g., cell-free bodily fluids, phagocytic cells, >2n phagocytic cells, circulating vesicles, and circulating diseased cells, or analytes isolated therefrom), may be referred to herein as a "combination sample." In some embodiments, the combination sample may comprise a lysate, or a fraction or portion of a lysate, of a population of phagocytic cells, a population of >2n phagocytic cells, a population of circulating vesicles, or a population of circulating diseased cells in place of the population of phagocytic cells, the population of >2n phagocytic cells, the population of circulating vesicles, or the population of circulating diseased cells, respectively. In some embodiments, the combination sample may comprise an analyte isolated from a lysate, or a fraction or portion of an analyte isolated from a lysate, of a population of phagocytic cells, a population of >2n phagocytic cells, a population of circulating vesicles, or a population of circulating diseased cells in place of the population of phagocytic cells, the population of >2n phagocytic cells, the population of circulating vesicles, or the population of circulating diseased cells, respectively. In some embodiments, the combination sample may comprise an analyte isolated from a cell-free bodily fluid, or a portion or fraction of an analyte isolated from a cell-free bodily fluid, rather than the cell-free bodily fluid itself. A control profile may be from phagocytic cells, phagocytic cells having a DNA content of 2n (=2n phagocytic cells), non-phagocytic cells, or control cells that are substantially free of cells affected by the disease or condition, taken from the same individual. Alternatively, the control profile may be a profile from a repository of markers of a disease or condition. In the context of this invention, "control cells that are substantially free of cells affected by the disease or condition" refers to a population of cells, as compared to circulating diseased cells, comprise significantly

fewer amounts of cells affected by the disease or condition. In some embodiments, "control cells that are substantially free of cells affected by the disease or condition" are cells that are at least 75%, 80%, 85%, 90%, 91%, 92%, 93%, 94%, 95%, 96%, 97%, 98%, 99%, 99.5%, or 100% free of cells affected by the disease or condition. In some embodiments, control cells that can be used in the methods of this invention are substantially free of fetal material (e.g., nucleic acids, proteins, and any analyte described herein), such as control cells that are at least 75%, 80%, 85%, 90%, 91%, 92%, 93%, 94%, 95%, 96%, 97%, 98%, 99%, 99.5%, or 100% free of fetal material.

[0032] This invention also provides methods for assessing the risk of developing a disease or condition, prognosing said disease, monitoring said disease progression or regression, assessing the efficacy of a treatment, or identifying a compound capable of ameliorating or treating said disease or condition.

[0033] In some embodiments, one or more components of whole blood that are substantially free of markers of a disease or condition may be isolated from a whole blood sample. In these embodiments, the one or more components of whole blood that are substantially free of markers may be used as a control sample (e.g., to determine a control profile), and the remaining portion of the whole blood sample may be used as an analytical sample (e.g., to determine an analytical profile). For example, a population of non-phagocytic cells may be isolated from a whole blood sample and used to determine a control profile, while the remaining portion of the whole blood sample is used to determine an analytical profile.

[0034] The methods of this invention can be used together with any known diagnostic methods, such as physical inspection, visual inspection, biopsy, scanning, histology, radiology, imaging, ultrasound, use of a commercial kit, genetic testing, immunological testing, analysis of bodily fluids, or monitoring neural activity.

[0035] Phagocytic cells that can be used in the methods of this invention include all types of cells that are capable of ingesting various types of substances (e.g., apoptotic cells, infectious agents, dead cells, viable cells, cell-free DNAs, cell-free RNAs, cell-free proteins). In some embodiments, the phagocytic cells are neutrophils, macrophages, monocytes, dendritic cells, foam cells, mast cells, eosinophils, or

keratinocytes. In some embodiments, the phagocytic cells can be a mixture of different types of phagocytic cells. In some embodiments, the phagocytic cells can be activated phagocytic cells., e.g., activated macrophages or neutrohils. In some embodiments, a phagocyte is a histiocyte, e.g., a Langerhans cell.

[0036] As used herein, ">2n phagocytic cells" refer to phagocytic cells that have a DNA content of greater than 2n, while "=2n phagocytic cells" refer to phagocytic cells that have a DNA content of 2n. According to the present invention, some phagocytic cells engulf live/dying/dead diseased cells (and sub-cellular fragments thereof) and/or cell-free disease-specific nucleic acids, proteins, carbohydrates and/or lipids present in bodily fluids. Such phagocytosis leads to the internalization of these disease markers into the phagocytic cell and, therefore, the DNA content of these phagocytic cells will become greater than 2n. By contrast, some phagocytic cells have not engulfed living/dying/dead diseased cells or fragments and/or cell-free disease-specific nucleic acids, proteins, lipids, and/or carbohydrates present in bodily fluids. The DNA contents of this group of phagocytic cells remain 2n. In some embodiments, the disease-specific markers (e.g., DNA with disease-specific mutations) can be expressed by the >2n phagocytic cells. For example, the mutated DNA of diseased cells is integrated into the normal DNA of the >2n phagocytic cells. The subsequent transcription of the "integrated" DNA of the >2n phagocytic cells into RNA and the translation of the latter into proteins produces a phenotype different from the phagocytic cells that have not phagocytosed the diseased cells (i.e., the =2n phagocytic cells). In other embodiments, the internalized disease-specific markers are not expressed by the >2n phagocytic cells. The markers may be translocated onto the membranes of the >2n phagocytic cells, or secreted out by the >2n phagocytic cells.

[0037] Circulating diseased cells that can be used in the methods of this invention include all types of circulating cells that may be affected by a disease or condition or infected by an infectious agent. A circulating cell refers to a cell present in the bodily fluid. A circulating cell may not necessarily circulate throughout the entire body or in the circulatory system. For example, a circulating cell may be present locally, such as in synovial fluid, or cerebrospinal fluid, or lymph fluid. A circulating diseased cell

may also be detached from a tissue or organ that has been affected by a disease or condition or infected by an infectious agent.

[0038] Cells that can be used as control cells in the methods of this invention includes all types of normal cells, or healthy cells, or cells that are substantially free of a disease or condition, or cells that are substantially free of an infectious agent. Control cells may be circulating cells or non-circulating cells (e.g., biopsied cells) that are representative of a normal or non-diseased state to which measurements on circulating diseased cells are compared to determine whether one or more diseased-associated marker is present in different levels between the circulating diseased cells and the control cells. The nature of the control cell may be a matter of design choice for a particular assay and may be derived or determined from normal tissue of the patient him- or herself.

[0039] In some embodiments, the circulating diseased cells are blood cells, tumor cells, lymphoma cells, fetal cells, apoptotic cells, epithelia cells, endothelial cells, stem cells, progenitor cells, mesenchymal cells, osteoblast cells, osteocytes, hematopoietic stem cells, foam cells, adipose cells, transcervical cells, circulating cardiocytes, circulating fibrocytes, circulating cancer stem cells, circulating myocytes, circulating cells from kidney, circulating cells from gastrointestinal tract, circulating cells from lung, circulating cells from reproductive organs, circulating cells from central nervous system, circulating hepatic cells, circulating cells from spleen, circulating cells from thymus, circulating cells from thyroid , circulating cells from an endocrine gland, circulating cells from parathyroid, circulating cells from pituitary, circulating cells from adrenal gland, circulating cells from islets of Langerhans, circulating cells from pancreas, circulating cells from hypothalamus, circulating cells from prostate tissues, circulating cells from breast tissues, circulating cells from circulating retinal cells, circulating ophthalmic cells, circulating auditory cells, circulating epidermal cells, or circulating cells from the urinary tract. In other embodiments, the circulating diseased cells can be a mixture of different types of circulating diseased cells.

[0040] The circulating diseased cells that can be used in the methods of this invention may be affected by various diseases or conditions. Exemplary diseases or conditions are a cardiovascular disease or condition, a kidney-associated disease or condition, a prenatal or pregnancy-related disease or condition, a neurological or neuropsychiatric disease or condition, an autoimmune or immune-related disease or condition, a cancer, an infectious disease or condition, a mitochondrial disorder, a respiratory-gastrointestinal tract disease or condition, a reproductive disease or condition, an ophthalmic disease or condition, a musculo-skeletal disease or condition, or a dermal disease or condition.

[0041] The circulating diseased cells that can be used in the methods of this invention may be infected by an infectious agent, such as a virus, bacteria, fungus, parasite, protozoan, infectious protein or microorganism.

[0042] In some embodiments, a cell useful in the methods of the invention (e.g., a circulating diseased cell, control cell not affected by the disease or condition, phagocytic cell, >2n phagocytic cell, =2n phagocytic cell, or non-phagocytic cell) is enucleated. Cells may be enucleated, for example, by using physical removal (e.g., via microneedle, optical tweezer, or aspiration), chemical treatments, photoablation, or ultraviolet irradiation.

[0043] As used herein, a "circulating vesicle" refers to a membrane-bound vesicle of cellular origin. The circulating vesicle may not necessarily circulate throughout the entire body or in the circulatory system. For example, the circulating vesicle may be present locally, such as in synovial fluid, or cerebrospinal fluid, or lymph fluid. In some embodiments, the circulating vesicles are selected from the group consisting of circulating microvesicles, apoptotic bodies, micro-particles, membrane-bound vesicles, multivesicular bodies, nanovesicles, microparticles, and ARRDC-1 mediated microvesicles (ARMM). In further embodiments, the circulating microvesicles are exosomes or urinary exosomes.

[0044] In some embodiments, a combination sample comprises two or more different components selected from the group consisting of: a cell-free bodily fluid isolated from a subject, a population of phagocytic cells isolated from the subject, a

population of >2n phagocytic cells isolated from the subject, a population of circulating vesicles isolated from the subject, and a population of circulating diseased cells isolated from the subject. In some embodiments, a combination sample comprises an analyte isolated from a cell-free bodily fluid isolated from the subject, an analyte isolated from a population of phagocytic cells isolated from the subject, an analyte isolated from a population of >2n phagocytic cells isolated from the subject, an analyte isolated from a population of circulating vesicles isolated from the subject, and an analyte isolated from a population of circulating diseased cells isolated from the subject.

[0045] As used herein, a "profile" of a marker of a disease or condition can broadly refer to any information concerning the marker. This information can be either qualitative (e.g., presence or absence) or quantitative (e.g., levels, copy numbers, or dosages). In some embodiments, a profile of a marker can indicate the absence of this marker. The profile can be a nucleic acid (e.g., DNA or RNA) profile, a protein profile, a lipid profile, a carbohydrate profile, a metabolite profile, or a combination thereof. A "marker" as used herein generally refers to an analyte which is differentially detectable in phagocytes and is indicative of the presence of a disease or condition. An analyte is differentially detectable if it can be distinguished quantitatively or qualitatively in phagocytes.

[0046] The methods of this invention can be applied to various diseases or conditions. Exemplary diseases or conditions are a cardiovascular disease or condition, a kidney-associated disease or condition, a prenatal or pregnancy-related disease or condition, a neurological or neuropsychiatric disease or condition, an autoimmune or immune-related disease or condition, a cancer, an infectious disease or condition, a pediatric disease, disorder, or condition, a mitochondrial disorder, a respiratory-gastrointestinal tract disease or condition, a reproductive disease or condition, an ophthalmic disease or condition, a musculo-skeletal disease or condition, or a dermal disease or condition.

[0047] As used herein, the term "cardiovascular disease or condition" refers to any condition that affects systems of heart or blood vessels (arteries and veins). Examples

of cardiovascular diseases include, but are not limited to myocardial infarction, coronary artery disease, percutaneous transluminal coronary angioplasty (PTCA), coronary artery bypass surgery (CABG), restenosis, peripheral arterial disease, stroke, abdominal aorta aneurysm, intracranial aneurysm, large artery atherosclerotic stroke, cardiogenic stroke, an early onset myocardial infarction, heart failure, pulmonary embolism, acute coronary syndrome (ACS), angina, cardiac hypertrophy, arteriosclerosis, myocarditis, pancarditis, endocarditis, hypertension, congestive heart failure, atherosclerosis, cerebrovascular disease, declining cardiac health, ischemic heart disease, pericarditis, cardiogenic shock, alcoholic cardiomyopathy, congenital heart disease, ischemic cardiomyopathy, hypertensive cardiomyopathy, valvular cardiomyopathy, inflammatory cardiomyopathy, cardiomyopathy secondary to a systemic metabolic disease, dilated cardiomyopathy, hypertrophic cardiomyopathy, arrhythmogenic right ventricular cardiomyopathy, restrictive cardiomyopathy, noncompaction cardiomyopathy, valvular heart disease, hypertensive heart disease, myocardial ischemic attack, unstable angina, myocardial rupture, cardiogenic shock, embolism, deep vein thrombosis, arrhythmia, arrhythmogenic right ventricular cardiomyopathy, diabetic cardiomyopathy, mitral regurgitation, mitral valve prolapse, peripheral vascular disease, artery disease, carotid artery disease, deep vein thrombosis, venous diseases, cerebrovascular disease, arterial aneurysm, left ventricular hypertrophy, hypertensive renal disease, hypertensive retinal disease, vasculitis, left main disease, arterial vascular disease, venous vascular disease, thrombosis of the microcirculation, a transient cerebrovascular accident, limb ischemia, aneurysm, thrombosis, superficial venous thrombosis, and deep venous thrombosis.

[0048] As used herein, the term "kidney-associated disease or condition" refers to any disease or condition that affects kidney or renal system. Examples of kidney-associated disease include, but are not limited to, chronic kidney diseases, primary kidney diseases, non-diabetic kidney diseases, glomerulonephritis, interstitial nephritis, diabetic kidney diseases, diabetic nephropathy, glomerulosclerosis, rapid progressive glomerulonephritis, renal fibrosis, Alport syndrome, IDDM nephritis, mesangial proliferative glomerulonephritis, membrano proliferative glomerulonephritis, crescentic glomerulonephritis, renal insterstitial fibrosis, focal

segmental glomerulosclerosis, membranous nephropathy, minimal change disease, pauci-immune rapid progressive glomerulonephritis, IgA nephropathy, polycystic kidney disease, Dent's disease, nephrocytinosis, Heymann nephritis, autosomal dominant (adult) polycystic kidney disease, autosomal recessive (childhood) polycystic kidney disease, acute kidney injury, nephrotic syndrome, renal ischemia, podocyte diseases or disorders, proteinuria, glomerular diseases, membranous glomerulonephritis, focal segmental glomerulonephritis, pre-eclampsia, eclampsia, kidney lesions, collagen vascular diseases, benign orthostatic (postural) proteinuria, IgM nephropathy, membranous nephropathy, sarcoidosis, diabetes mellitus, kidney damage due to drugs, Fabry's disease, aminoaciduria, Fanconi syndrome, hypertensive nephrosclerosis, interstitial nephritis, Sickle cell disease, hemoglobinuria, myoglobinuria, Wegener's Granulomatosis, Glycogen Storage Disease Type 1, chronic kidney disease, chronic renal failure, low Glomerular Filtration Rate (GFR), nephroangiosclerosis, lupus nephritis, ANCA-positive pauci-immune crescentic glomerulonephritis, chronic allograft nephropathy, nephrotoxicity, renal toxicity, kidney necrosis, kidney damage, glomerular and tubular injury, kidney dysfunction, nephritic syndrome, acute renal failure, chronic renal failure, proximal tubal dysfunction, acute kidney transplant rejection, chronic kidney transplant refection, non IgA mesangioproliferative glomerulonephritis, postinfectious glomerulonephritis, vasculitides with renal involvement of any kind, any hereditary renal disease, any interstitial nephritis, renal transplant failure, kidney cancer, kidney disease associated with other conditions (e.g., hypertension, diabetes, and autoimmune disease), Dent's disease, nephrocytinosis, Heymann nephritis, a primary kidney disease, a collapsing glomerulopathy, a dense deposit disease, a cryoglobulinemia-associated glomerulonephritis, an Henoch-Schónlein disease, a postinfectious glomerulonephritis, a bacterial endocarditis, a microscopic polyangitis, a Churg-Strauss syndrome, an anti-GBM-antibidy mediated glomerulonephritis, amyloidosis, a monoclonal immunoglobulin deposition disease, a fibrillary glomerulonephritis, an immunotactoid glomerulopathy, ischemic tubular injury, a medication-induced tubulo-interstitial nephritis, a toxic tubulo-interstitial nephritis, an infectious tubulo-interstitial nephritis, a bacterial pyelonephritis, a viral infectious tubulo-interstitial nephritis which results from a polyomavirus infection or an HIV infection, a

metabolic-induced tubulo-interstitial disease, a mixed connective disease, a cast nephropathy, a crystal nephropathy which may results from urate or oxalate or drug-induced crystal deposition, an acute cellular tubulo-interstitial allograft rejection, a tumoral infiltrative disease which results from a lymphoma or a post-transplant lymphoproliferative disease, an obstructive disease of the kidney, vascular disease, a thrombotic microangiopathy, a nephroangiosclerosis, an atheroembolic disease, a mixed connective tissue disease, a polyarteritis nodosa, a calcineurin-inhibitor induced- vascular disease, an acute cellular vascular allograft rejection, an acute humoral allograft rejection, early renal function decline (ERFD), end stage renal disease (ESRD), renal vein thrombosis, acute tubular necrosis, acute interstitial nephritis, established chronic kidney disease, renal artery stenosis, ischemic nephropathy, uremia, drug and toxin-induced chronic tubulointerstitial nephritis, reflux nephropathy, kidney stones, Goodpasture's syndrome, and hydronephrosis.

[0049] As used herein, the term "prenatal or pregnancy-related disease or condition" refers to any disease, disorder, or condition affecting a pregnant woman, embryo, or fetus. Prenatal or pregancy-related conditions can also refer to any disease, disorder, or condition that is associated with or arises, either directly or indirectly, as a result of pregnancy. These diseases or conditions can include any and all birth defects, congenital conditions, or hereditary diseases or conditions. Examples of prenatal or pregnancy-related diseases include, but are not limited to, Rhesus disease, hemolytic disease of the newborn, beta-thalassemia, sex determination, determination of pregnancy, a hereditary Mendelian genetic disorder, chromosomal aberrations, a fetal chromosomal aneuploidy, fetal chromosomal trisomy, fetal chromosomal monosomy, trisomy 8, trisomy 13 (Patau Syndrom), trisomy 16, trisomy 18 (Edwards syndrome), trisomy 21 (Down syndrome), X-chromosome linked disorders, trisomy X (XXX syndrome), monosomy X (Turner syndrome), XXY syndrome, XYY syndrome, XYY syndrome, XXXY syndrome, XXYY syndrome, XYYY syndrome, XXXXX syndrome, XXXXY syndrome, XXXYY syndrome, XXYYY syndrome, Fragile X Syndrome, fetal growth restriction, cystic fibrosis, a hemoglobinopathy, fetal death, fetal alcohol syndrome, sickle cell anemia, hemophilia, Klinefelter syndrome, dup(17)(p11.2p1.2) syndrome, endometriosis, Pelizaeus-Merzbacher disease, dup(22)(q11.2q11.2) syndrome, cat eye syndrome, cri-du-chat syndrome, Wolf-

Hirschhorn syndrome, Williams-Beuren syndrome, Charcot-Marie-Tooth disease, neuropathy with liability to pressure palsies, Smith-Magenis syndrome, neurofibromatosis, Alagille syndrome, Velocardiofacial syndrome, DiGeorge syndrome, steroid sulfatase deficiency, Prader-Willi syndrome, Kallmann syndrome, microphthalmia with linear skin defects, adrenal hypoplasia, glycerol kinase deficiency, Pelizaeus-Merzbacher disease, testis-determining factor on Y, azospermia (factor a), azospermia (factor b), azospermia (factor c), 1p36 deletion, phenylketonuria, Tay-Sachs disease, adrenal hyperplasia, Fanconi anemia, spinal muscular atrophy, Duchenne's muscular dystrophy, Huntington's disease, myotonic dystrophy, Robertsonian translocation, Angelman syndrome, tuberous sclerosis, ataxia telangieltasia, open spina bifida, neural tube defects, ventral wall defects, small-for-gestational-age, congenital cytomegalovirus, achondroplasia, Marfan's syndrome, congenital hypothyroidism, congenital toxoplasmosis, biotinidase deficiency, galactosemia, maple syrup urine disease, homocystinuria, medium-chain acyl Co-A dehydrogenase deficiency, structural birth defects, heart defects, abnormal limbs, club foot, anencephaly, arhinencephaly/holoprosencephaly, hydrocephaly, anophthalmos/microphthalmos, anotia/microtia, transposition of great vessels, tetralogy of Fallot, hypoplastic left heart syndrome, coarctation of aorta, cleft palate without cleft lip, cleft lip with or without cleft palate, oesophageal atresia/stenosis with or without fistula, small intestine atresia/stenosis, anorectal atresia/stenosis, hypospadias, indeterminate sex, renal agenesis, cystic kidney, preaxial polydactyly, limb reduction defects, diaphragmatic hernia, blindness, cataracts, visual problems, hearing loss, deafness, X-linked adrenoleukodystrophy, Rett syndrome, lysosomal disorders, cerebral palsy, autism, aglossia, albinism, ocular albinism, oculocutaneous albinism, gestational diabetes, Arnold-Chiari malformation, CHARGE syndrome, congenital diaphragmatic hernia, brachydactlia, aniridia, cleft foot and hand, heterochromia, Dwarnian ear, Ehlers Danlos syndrome, epidermolysis bullosa, Gorham's disease, Hashimoto's syndrome, hydrops fetalis, hypotonia, Klippel-Feil syndrome, muscular dystrophy, osteogenesis imperfecta, progeria, Smith Lemli Opitz symdrom, chromatelopsia, X-linked lymphoproliferative disease, omphalocele, gastroschisis, pre-eclampsia, eclampsia, pre-term labor, premature birth, miscarriage,

delayed intrauterine growth, ectopic pregnancy, hyperemesis gravidarum, morning sickness, or likelihood for successful induction of labor.

[0050] As used herein, the term "a neurological or neuropsychiatric disease or condition" refers to any disease or condition that affects nervous systems. Examples of neurological or neuropsychiatric diseases or conditions include, but are not limited to, head trauma, stroke, stroke, ischemic stroke, hemorrhagic stroke, subarachnoid hemorrhage, intra cranial hemorrhage, transient ischemic attack, vascular dementia, corticobasal ganglionic degeneration, encephalitis, epilepsy, Landau-Kleffner syndrome, hydrocephalus, pseudotumor cerebri, thalamic diseases, meningitis, myelitis, movement disorders, essential tremor, spinal cord diseases, syringomyelia, Alzheimer's disease (early onset), Alzheimer's disease (late onset), multi-infarct dementia, Pick's disease, Huntingdon's disease, Parkinson's disease, Parkinson syndromes, dementia, frontotemporal dementia, corticobasal degeneration, multiple system atrophy, progressive supranuclear palsy, Lewy body disease, Creutzfeldt-Jakob disease, Dandy-Walker syndrome, Friedreich ataxia, Machado-Joseph disease, migraine, schizophrenia, mood disorders and depression. dementia with lewy bodies (DLB), frontotemporal dementia (FTD), various forms of vascular dementia (VD), subcortical vascular dementia (Binswanger's disease), autism, developmental retardations, motor neuron diseases, amyotrophic lateral sclerosis (ALS), neuronal or brain damage, hypoxia of the brain, cerebral palsy (CP), memory disorders, movement disorders, corticalbasal ganglionic degeneration, forms of multiple system atrophy, stroke-related disorders, cerebrovascular accidents, post-irradiation encephalopathy with seizures, vascular Parkinsonism, thalamic cerebrovascular accidents, chronic inflammatory demyelinating polyneuropathy, alcohol related dementia, semantic dementia, ataxia, atypical Parkinsonism, dystonia, progressive supranuclear palsy, essential tremor, mild cognitive impairment, amyotrophic lateral sclerosis, multiple sclerosis, neuropathies, Pick's disease, congophilic amyloid angiopathy, Creutzfeldt-Jakob Disease, AIDS dementia complex, depression, anxiety disorder, phobia, Bell's Palsy, epilepsy, encephalitis, neuromuscular disorders, neurooncological disorders, brain tumors, neurovascular disorders, neuroimmunological disorders, neurootological disease, neurotrauma including spinal cord injury, pain including neuropathic pain, pediatric neurological and

neuropsychiatric disorders, sleep disorders, Tourette syndrome, corticalbasal ganglionic degeneration, alcohol related dementia, semantic dementia, Alzheimer's disease combined with multi-infarct dementia, Alzheimer's disease combined with Lewy body dementia, Parkinson's disease combined with Lewy body dementia, Alzheimer's and Parkinson's disease combined with Lewy body dementia, frontotemporal dementia combined with chronic inflammatory demyelinating polyneuropathy, attention deficit hyperactivity disorder, schizophrenia, obsessive-compulsive disorder, mental retardation, autistic spectrum disorders, opsoclonus-myoclonus syndrome (OMS) seizures, articulation disorder, learning disabilities (i.e., reading or arithmetic), verbal or performance aptitude deficits, attention deficit disorder, amyloid diseases, prion diseases, Tauopathies, Alpha-Synucleinopathies, addictive states such as those caused by at least one of: cocaine, nicotine, alcohol, food, ecstasy, kat, caffeine, opium, heroin, marijuana, amphetamine, methamphetamine or gambling, and Fabry's disease.

[0051] As used herein, the term "an autoimmune or immune-related disease or condition" refers to any disease or condition that affects the function of immune systems. Examples of autoimmune or immune-related diseases or conditions include, but are not limited to, antiphospholipid syndrome, systemic lupus erythematosus, rheumatoid arthritis, autoimmune vasculitis, celiac disease, autoimmune thyroiditis, post-transfusion immunization, maternal-fetal incompatibility, transfusion reactions, immunological deficiency such IgA deficiency, common variable immunodeficiency, drug-induced lupus, diabetes mellitus, Type I diabetes, Type II diabetes, juvenile onset diabetes, juvenile rheumatoid arthritis, psoriatic arthritis, multiple sclerosis, immunodeficiency, allergies, asthma, psoriasis, atopic dermatitis, allergic contact dermatitis, chronic skin diseases, amyotrophic lateral sclerosis, chemotherapy-induced injury, graft-vs-host diseases, bone marrow transplant rejection, Ankylosing spondylitis, atopic eczema, Pemphigus, Behcet's disease, chronic fatigue syndrome fibromyalgia, chemotherapy-induced injury, myasthenia gravis, glomerulonephritis, allergic retinitis, systemic sclerosis, subacute cutaneous lupus erythematosus, cutaneous lupus erythematosus including chilblain lupus erythematosus, Sjogren's syndrome, autoimmune nephritis, autoimmune vasculitis, autoimmune hepatitis, autoimmune carditis, autoimmune encephalitis, autoimmune mediated hematological

diseases, lc-SSc (limited cutaneous form of scleroderma), dc-SSc (diffused cutaneous form of scleroderma), autoimmune thyroiditis (AT), Grave's disease (GD), myasthenia gravis, multiple sclerosis (MS), ankylosing spondylitis. transplant rejection, immune aging, rheumatic/autoimmune diseases, mixed connective tissue disease, spondyloarthropathy, psoriasis, psoriatic arthritis, myositis, scleroderma, dermatomyositis, autoimmune vasculitis, mixed connective tissue disease, idiopathic thrombocytopenic purpura, Crohn's disease, human adjuvant disease, osteoarthritis, juvenile chronic arthritis, a spondyloarthropathy, an idiopathic inflammatory myopathy, systemic vasculitis, sarcoidosis, autoimmune hemolytic anemia, autoimmune thrombocytopenia, thyroiditis, immune-mediated renal disease, a demyelinating disease of the central or peripheral nervous system, idiopathic demyelinating polyneuropathy, Guillain-Barre syndrome, a chronic inflammatory demyelinating polyneuropathy, a hepatobiliary disease, infectious or autoimmune chronic active hepatitis, primary biliary cirrhosis, granulomatous hepatitis, sclerosing cholangitis, inflammatory bowel disease, gluten-sensitive enteropathy, Whipple's disease, an autoimmune or immune-mediated skin disease, a bullous skin disease, erythema multiforme, allergic rhinitis, atopic dermatitis, food hypersensitivity, urticaria, an immunologic disease of the lung, eosinophilic pneumonias, idiopathic pulmonary fibrosis, hypersensitivity pneumonitis, a transplantation associated disease, graft rejection or graft-versus-host-disease, psoriatic arthritis, psoriasis, dermatitis, polymyositis/dermatomyositis, toxic epidermal necrolysis, systemic scleroderma and sclerosis, responses associated with inflammatory bowel disease, Crohn's disease, ulcerative colitis, respiratory distress syndrome, adult respiratory distress syndrome (ARDS), meningitis, encephalitis, uveitis, colitis, glomerulonephritis, allergic conditions, eczema, asthma, conditions involving infiltration of T cells and chronic inflammatory responses, atherosclerosis, autoimmune myocarditis, leukocyte adhesion deficiency, allergic encephalomyelitis, immune responses associated with acute and delayed hypersensitivity mediated by cytokines and T-lymphocytes, tuberculosis, sarcoidosis, granulomatosis including Wegener's granulomatosis, agranulocytosis, vasculitis (including ANCA), aplastic anemia, Diamond Blackfan anemia, immune hemolytic anemia including autoimmune hemolytic anemia (AIHA), pernicious anemia, pure red cell aplasia (PRCA), Factor VIII deficiency, hemophilia

A, autoimmune neutropenia, pancytopenia, leukopenia, diseases involving leukocyte diapedesis, central nervous system (CNS) inflammatory disorders, multiple organ injury syndrome, mysathenia gravis, antigen-antibody complex mediated diseases, anti-glomerular basement membrane disease, anti-phospholipid antibody syndrome, allergic neuritis, Bechet disease, Castleman's syndrome, Goodpasture's syndrome, Lambert-Eaton Myasthenic Syndrome, Reynaud's syndrome, Sjorgen's syndrome, Stevens-Johnson syndrome, pemphigoid bullous, pemphigus, autoimmune polyendocrinopathies, Reiter's disease, stiff-man syndrome, giant cell arteritis, immune complex nephritis, IgA nephropathy, IgM polyneuropathies or IgM mediated neuropathy, idiopathic thrombocytopenic purpura (ITP), thrombotic throbocytopenic purpura (TTP), autoimmune thrombocytopenia, autoimmune disease of the testis and ovary including autoimmune orchitis and oophoritis, primary hypothyroidism, autoimmune endocrine diseases including autoimmune thyroiditis, chronic thyroiditis (Hashimoto's Thyroiditis), subacute thyroiditis, idiopathic hypothyroidism, Addison's disease, Grave's disease, autoimmune polyglandular syndromes (or polyglandular endocrinopathy syndromes), Sheehan's syndrome, autoimmune hepatitis, lymphoid interstitial pneumonitis (HIV), bronchiolitis obliterans (non-transplant) vs NSIP, Guillain-Barre' Syndrome, large vessel vasculitis (including polymyalgia rheumatica and giant cell (Takayasu's) arteritis), medium vessel vasculitis (including Kawasaki's disease and polyarteritis nodosa), ankylosing spondylitis, Berger's disease (IgA nephropathy), rapidly progressive glomerulonephritis, primary biliary cirrhosis, Celiac sprue (gluten enteropathy), cryoglobulinemia, and amyotrophic lateral sclerosis (ALS).

[0052] As used herein, the term "cancer" refers to various types of malignant neoplasms, most of which can invade surrounding tissues, and may metastasize to different sites (see, for example, PDR Medical Dictionary, 1st edition (1995), incorporated herein by reference in its entirety for all purposes). The terms "neoplasm" and "tumor" refer to an abnormal tissue that grows by cellular proliferation more rapidly than normal and continues to grow after the stimuli that initiated proliferation is removed. Such abnormal tissue shows partial or complete lack of structural organization and functional coordination with the normal tissue which may be either benign (i.e., benign tumor) or malignant (i.e., malignant tumor).

Examples of general categories of cancer include, but are not limited to, carcinomas (i.e., malignant tumors derived from epithelial cells such as, for example, common forms of breast, prostate, lung and colon cancer), sarcomas (i.e., malignant tumors derived from connective tissue or mesenchymal cells), lymphomas (i.e., malignancies derived from hematopoietic cells), leukemias (i.e., malignancies derived from hematopoietic cells), germ cell tumors (i.e., tumors derived from totipotent cells. In adults most often found in the testicle or ovary; in fetuses, babies and young children, most often found on the body midline, particularly at the tip of the tailbone), blastic tumors (i.e., a typically malignant tumor which resembles an immature or embryonic tissue) and the like. Examples of the types of neoplasms intended to be encompassed by the present invention include but are not limited to those neoplasms associated with cancers of neural tissue, blood forming tissue, breast, skin, bone, prostate, ovaries, uterus, cervix, liver, lung, brain, larynx, gallbladder, pancreas, rectum, parathyroid, thyroid, adrenal gland, immune system, head and neck, colon, stomach, bronchi, and/or kidneys.

[0053] As used herein, the term "an infectious disease or condition" refers to any disease or condition that results from an infectious agent. Infectious agents include, but are not limited to bacteria, viruses, fungi, protozoa, infectious proteins, parasitic microbes, and other parasites. Examples of infectious diseases or conditions include, but are not limited to, bacterial infections, viral infections, fungal infections, protozoan infections, parasitic infections, hepatitis (e.g., hepatitis A, B, C, D, and E), herpes, influenza, human papillomavirus (HPV) infection, AIDS, anthrax, pneumonia (bacterial or viral), cellulitis, human parainfluenza, the common cold, Legionnaires' disease (Legionellosis), cholera, Creutzfeldt-Jakob disease (CJD), variant Creutzfeldt-Jakob disease (vCJD), fatal familial insomnia (FFI), Gerstmann-Sträussler-Scheinker (GSS) syndrome, Chlamydia, chicken pox, ebola hemorrhagic fever, Dengue fever, giardiasis, Lyme disease, malaria, measles, mumps, rubella, pertussis, gonorrhea, staphylococcal infection, streptococcal infection, pneumococcal infection, rabies, helicobacter pylori infection, respiratory syncitial virus infection, Rocky Mountain spotted fever, SARS, sepsis, tuberculosis, and West Nile fever.

[0054] Viruses include, but are not limited to, DNA or RNA animal viruses. As used herein, RNA viruses include, but are not limited to, virus families such as Picornaviridae (e.g., polioviruses), Reoviridae (e.g., rotaviruses), Togaviridae (e.g., encephalitis viruses, yellow fever virus, rubella virus), Orthomyxoviridae (e.g., influenza viruses), Paramyxoviridae (e.g., respiratory syncytial virus, measles virus, mumps virus, parainfluenza virus), Rhabdoviridae (e.g., rabies virus), Coronaviridae, Bunyaviridae, Flaviviridae, Filoviridae, Arenaviridae, Bunyaviridae and Retroviridae (e.g., human T cell lymphotropic viruses (HTLV), human immunodeficiency viruses (HIV)). As used herein, DNA viruses include, but are not limited to, virus families such as Papovaviridae (e.g., papilloma viruses), Adenoviridae (e.g., adenovirus), Herpesviridae (e.g., herpes simplex viruses), and Poxviridae (e.g., variola viruses).

[0055] Bacteria include, but are not limited to, gram positive bacteria, gram negative bacteria, acid-fast bacteria and the like. Gram positive bacteria include, but are not limited to, Actinomedurae, Actinomyces israelii, Bacillus anthracis, Bacillus cereus, Clostridium botulinum, Clostridium difficile, Clostridium perfringens, Clostridium tetani, Corynebacterium, Enterococcus faecalis, Listeria monocytogenes, Nocardia, Propionibacterium acnes, Staphylococcus aureus, Staphylococcus epiderm, Streptococcus mutans, Streptococcus pneumoniae and the like. Gram negative bacteria include, but are not limited to, Afipia felis, Bacteriodes, Bartonella bacilliformis, Bortadella pertussis, Borrelia burgdorferi, Borrelia recurrentis, Brucella, Calymmatobacterium granulomatis, Campylobacter, Escherichia coli, Francisella tularensis, Gardnerella vaginalis, Haemophilius aegyptius, Haemophilius ducreyi, Haemophilius influenziae, Heliobacter pylori, Legionella pneumophila, Leptospira interrogans, Neisseria meningitidia, Porphyromonas gingivalis, Providencia sturti, Pseudomonas aeruginosa, Salmonella enteridis, Salmonella typhi, Serratia marcescens, Shigella boydii, Streptobacillus moniliformis, Streptococcus pyogenes, Treponema pallidum, Vibrio cholerae, Yersinia enterocolitica, Yersinia pestis and the like. As used herein, acid-fast bacteria include, but are not limited to, Myobacterium avium, Myobacterium leprae, Myobacterium tuberculosis and the like. Other bacteria not falling into the other three categories include, but are not limited to, Bartonella henseiae, Chlamydia psittaci, Chlamydia trachomatis, Coxiella burnetii, Mycoplasma pneumoniae, Rickettsia akari, Rickettsia prowazekii, Rickettsia rickettsii, Rickettsia

tsutsugamushi, Rickettsia typhi, Ureaplasma urealyticum, Diplococcus pneumoniae, Ehrlichia chafensis, Enterococcus faecium, Meningococci and the like.

[0056] As used herein, fungi include, but are not limited to, Aspergilli, Candidae, Candida albicans, Coccidioides immitis, Cryptococci, and combinations thereof.

[0057] As used herein, parasitic microbes include, but are not limited to, Balantidium coli, Cryptosporidium parvum, Cyclospora cayatanensis, Encephalitozoa, Entamoeba histolytica, Enterocytozoon bieneusi, Giardia lamblia, Leishmaniae, Plasmodii, Toxoplasma gondii, Trypanosomae, trapezoidal amoeba and the like. Other parasites include worms (e.g., helminthes), particularly parasitic worms including, but not limited to, Nematoda (roundworms, e.g., whipworms, hookworms, pinworms, ascarids, filarids and the like), and Cestoda (e.g., tapeworms).

[0058] As used herein, infectious proteins include prions (e.g., PrP^{Sc} forms, the CJD prion, the vCJD prion, the FFI prion, and the GSS prion).

[0059] As used herein, the term "a pediatric disease, disorder, or condition" refers to any disease or disorder that affects infants or children or that begins during development or childhood. Examples of pediatric diseases, disorders, and conditions include, but are not limited to, autism, Kawasaki's disease, congenital deafness, pediatric cancers, Type I diabetes, congenital heart defects, tetralogy of Fallot, Duchenne Muscular Dystrophy, osteogenesis importfect, Krabe disease, Pompe disease, Gaucher disease, Fabry disease, Wolff-Parkinson-White syndrome, Hirschsprung's disease, Crohn's disease, Eagle-Barrett Syndrome, cystic fibrosis, irritable bowel syndrome, and cerebral palsy. Examples of pediatric conditions also include genetic attributes of the developing fetus. For example, a pediatric conditions include, but are not limited to, intelligence, eye color, hair color, and muscle type.

[0060] As used herein, "treating" a disease or condition refers to taking steps to obtain beneficial or desired results, including clinical results. Beneficial or desired clinical results include, but are not limited to, alleviation or amelioration of one or more symptoms associated with diseases or conditions.

[0061] As used herein, "administering" or "administration of" a compound or an agent to a subject can be carried out using one of a variety of methods known to those skilled in the art. For example, a compound or an agent can be administered, intravenously, arterially, intradermally, intramuscularly, intraperitonealy, intravenously, subcutaneously, ocularly, sublingually, orally (by ingestion), intranasally (by inhalation), intraspinally, intracerebrally, and transdermally (by absorbtion, e.g., through a skin duct). A compound or agent can also appropriately be introduced by rechargeable or biodegradable polymeric devices or other devices, e.g., patches and pumps, or formulations, which provide for the extended, slow, or controlled release of the compound or agent. Administering can also be performed, for example, once, a plurality of times, and/or over one or more extended periods. In some aspects, the administration includes both direct administration, including self-administration, and indirect administration, including the act of prescribing a drug. For example, as used herein, a physician who instructs a patient to self-administer a drug, or to have the drug administered by another and/or who provides a patient with a prescription for a drug is administering the drug to the patient. In some embodiments, a compound or an agent is administered orally, e.g., to a subject by ingestion, or intravenously, e.g., to a subject by injection. In some embodiments, the orally administered compound or agent is in an extended release or slow release formulation, or administered using a device for such slow or extended release.

[0062] In certain embodiments, markers used in the methods of invention are up-regulated or activated in the combination sample compared to the phagocytic cells, =2n phagocytic cells, non-phagocytic cells, or control cells that are substantially free of cells affected by the disease or condition. Different diseases or conditions can be associated with either up-regulation (or activation) or down-regulation (or inhibition) of different markers. As used herein, "up-regulation or up-regulated" can refer to an increase in expression levels (e.g., gene expression or protein expression), gene copy numbers, gene dosages, and other qualitative or quantitative detectable state of the markers. Similarly, "down-regulation or down-regulated" can refer to an increase in expression levels, gene copy numbers, gene dosages, and other qualitative or quantitative detectable state of the markers. In certain embodiments, markers used in the methods of invention are down-regulated or inhibited in the combination sample

compared to the phagocytic cells, =2n phagocytic cells, non-phagocytic cells, or control cells that are substantially free of cells affected by the disease or condition. Different diseases or conditions can be associated with either up-regulation (or activation) or down-regulation (or inhibition) of different markers. As used herein, "activation or activated" can refer to an active state of the marker, e.g., a phosphorylation state, a DNA methylation state, or a DNA acetylation state. Similarly, "inhibition or inhibited" can refer to a repressed state or an inactivated state of the marker, e.g., a de-phosphorylation state, a ubiquitination state, a DNA de-methylation state.

[0063] In certain embodiments, methods of this invention can comprise extracting or enriching markers from cell-free bodily fluids. Any known extraction and enrichment methods can be used herein. In certain embodiments, methods of this invention also comprise at least one of the following steps before determination of various profiles: i) lysing the =2n phagocytic, non-phagocytic, or control cells; and ii) extracting cellular contents from the lysed cells. Any known cell lysis and extraction methods can be used herein. In certain embodiments, at least one or more markers of a disease or condition are present in the combination sample. In certain embodiments, there is no marker present in the cellular contents of the non-phagocytic, =2n phagocytic, or control cells.

[0064] In certain embodiments, methods of this invention also comprise at least one of the following steps before determination of various profiles: i) lysing the phagocytic, >2n phagocytic cells, or =2n phagocytic cells; and ii) extracting cellular contents from the lysed phagocytic, >2n phagocytic, or =2n phagocytic cells. In certain embodiments, the cellular contents of the phagocytic or >2n phagocytic cells comprise various types of materials that they have engulfed, such as viable diseased cells, dead diseased cells, apoptotic diseased cells, circulating tumor cells, infectious agents, fetal cells, trophoblasts, or fragments thereof. In certain embodiments, at least one or more markers of a disease or condition are present in the cellular contents of the phagocytic or >2n phagocytic cells.

[0065] In certain embodiments, methods of this invention also comprise at least one of the following steps before determination of various profiles: i) lysing the circulating vesicles; and ii) extracting the contents from the lysed circulating vesicles. In certain embodiments, the contents of the circulating vesicles comprise various types of materials, such as proteins and nucleic acids. In certain embodiments, at least one or more markers of a disease or condition are present in the contents of the circulating vesicles.

[0066] In certain embodiments, methods of this invention also comprise at least one of the following steps before determination of various profiles: i) lysing the circulating diseased cells; and ii) extracting the contents from the lysed circulating diseased cells. In certain embodiments, at least one or more markers of a disease or condition are present in the contents of the circulating diseased cells.

[0067] In certain embodiments, methods of this invention further comprise comparing the identified difference of the disease or condition-specific markers to a repository of at least one markers known in the art. Such comparison can further confirm the presence of the disease or condition. In some embodiments, the repository of the known markers can be obtained by data mining. The term "data mining", as used herein, refers to a process of finding new data patterns, relations, or correlations derived from the known data of the databases and of extracting practicable information in the future. Typically a computer-based system can be trained on data to perform the data mining, e.g., to classify the input data and then subsequently used with new input data to make decisions based on the training data. These systems include, but are not limited, expert systems, fuzzy logic, non-linear regression analysis, multivariate analysis, decision tree classifiers, and Bayesian belief networks.

[0068] In certain embodiments, the phagocytic cells, >2n phagocytic cells, circulating vesicles, circulating diseased cells, control cells, and =2n phagocytic cells are isolated from a bodily fluid sample, tissues, or cells. Exemplary bodily fluid samples can be whole blood, urine, stool, saliva, lymph fluid, cerebrospinal fluid, synovial fluid, cystic fluid, ascites, pleural effusion, fluid obtained from a pregnant

woman in the first trimester, fluid obtained from a pregnant woman in the second trimester, fluid obtained from a pregnant woman in the third trimester, maternal blood, amniotic fluid, chorionic villus sample, fluid from a preimplantation embryo, maternal urine, maternal saliva, placental sample, fetal blood, lavage and cervical vaginal fluid, interstitial fluid, buccal swab sample, sputum, bronchial lavage, Pap smear sample, or ocular fluid. In some embodiments, the phagocytic cells, >2n phagocytic cells, and =2n phagocytic cells are isolated from white blood cells. In certain embodiments, the >2n phagocytic cells and the =2n phagocytic cells are separated from a population of phagocytic cells.

[0069] In certain embodiments, the cell-free bodily fluid comes from a bodily fluid sample. Exemplary bodily fluid samples can be whole blood, urine, stool, saliva, lymph fluid, cerebrospinal fluid, synovial fluid, cystic fluid, ascites, pleural effusion, fluid obtained from a pregnant woman in the first trimester, fluid obtained from a pregnant woman in the second trimester, fluid obtained from a pregnant woman in the third trimester, maternal blood, amniotic fluid, chorionic villus sample, fluid from a preimplantation embryo, maternal urine, maternal saliva, placental sample, fetal blood, lavage and cervical vaginal fluid, interstitial fluid, buccal swab sample, sputum, bronchial lavage, Pap smear sample, or ocular fluid. In some embodiments, the cell-free bodily fluids are obtained by separating cells from the bodily fluid sample by methods known in the art, such as extraction, centrifugation, and filtration.

[0070] In some embodiments, a component for use in a combination sample may be obtained by removing cells from a bodily fluid. In some embodiments, a component for use in a combination sample may be obtained by destroying (e.g., lysing) cells in a bodily fluid. These embodiments may be taken in combination, for example, by removing some populations of cells and destroying other populations of cells. In some embodiments, a sample of whole blood may be used to create a combination sample, for example, by removing red blood cells, serum, and T cells, and using the remainder as the combination sample.

[0071] In certain embodiments, tissue or fluid samples including cells having a DNA content of 2n are obtained post separation (e.g., via centrifugation) of non-

cellular fraction of fluids obtained by puncture of a vein or artery followed by the withdrawal of blood, tissue biopsies, bronchoalveolar lavage, nasal lavage, eye lavage, peritoneal cavity lavage, vaginal lavage, bladder lavage, rectal lavage, fine needle aspiration of spinal fluid, synovial fluid aspiration, and the like. Cell free bodily fluids are obtained post separation (e.g., via centrifugation) of cellular fraction of fluids obtained by puncture of a vein or artery followed by the withdrawal of blood, tissue biopsies, bronchoalveolar lavage, nasal lavage, eye lavage, peritoneal cavity lavage, vaginal lavage, bladder lavage, rectal lavage, fine needle aspiration of spinal fluid, synovial fluid aspiration, and the like.

[0072] In the methods of this invention, cell separation/isolation/purification methods are used to isolate populations of cells from bodily fluid sample, cells, or tissues of a subject. A skilled worker can use any known cell separation/isolation/purification techniques to isolate phagocytic cells, >2n phagocytic cells, diseased cells, control cells, and =2n phagocytic cells from a bodily fluid, or to separate >2n phagocytic cells from =2n phagocytic cells. Exemplary techniques include, but are not limited to, using antibodies, flow cytometry, fluorescence activated cell sorting, filtration, gradient-based centrifugation, elution, microfluidics, magnetic separation technique, fluorescent-magnetic separation technique, nanostructure, quantum dots, high throughput microscope-based platform, or a combination thereof.

[0073] In the methods of this invention, cell separation/isolation/purification methods are used to isolate populations of circulating diseased cells from bodily fluid sample, cells, or tissues of a subject. Circulating diseased cells may be rare or in low quantity in a bodily fluid. Therefore, enrichment techniques (e.g., magnetic enrichment) may be used to enrich circulating diseased cells before the isolation. A skilled worker can use any known cell separation/isolation/purification techniques to isolate circulating diseased cells from a bodily fluid. Exemplary techniques include, but are not limited to, using antibodies, flow cytometry, fluorescence activated cell sorting, filtration, gradient-based centrifugation, elution, microfluidics, magnetic separation technique, fluorescent-magnetic separation technique, nanostructure, quantum dots, high throughput microscope-based platform, micro-fluidic technique,

fiber-optic array–scanning technique, laser-scanning cytometry technique, multiphoton intravital flow cytometry, photoacoustic flowmetry, nanoparticles targeting cell surface antigens, staining circulating diseased cells with detectable secreted products, or a combination thereof. Circulating diseased cells may have different physical properties compared to normal circulating cells, such as difference in size, density, charge, migratory properties, and some properties of specific cell types (e.g., melanocytic granules in circulating melanoma cells). A skilled worker can use any known cell separation/isolation/purification techniques based on such different properties to isolate circulating diseased cells. For example, differences in buoyant density may be used to separate circulating diseased cells (e.g., circulating tumor cells) from normal blood cells through gradient centrifugation. Filtration-based approaches may be used isolate circulating diseased cells (e.g., circulating tumor cells) based on their increased sizes compared to normal circulating cells. Antibody-based isolation approaches may be used to capture circulating diseased cells, which express epithelia cell surface markers that are absent from normal circulating blood cells. For example, conjugation of antibodies against epithelial cell adhesion molecule (EpCAM) to magnetic beads, followed by purification of captured cells through a magnetic field, may be used to enrich circulating tumor cells from the blood of patients with cancers of the breast, prostate, and colon. In certain embodiments, circulating diseased cells (e.g., transcervical cells) may be collected by the RareCellect™ device (Genetic Technologies) or similar devices.

[0074] In certain embodiments, the phagocytic cells, >2n phagocytic cells, diseased cells, control cells, and =2n phagocytic cells are isolated by using a product secreted by the cells. In certain embodiments, the phagocytic cells, >2n phagocytic cells, diseased cells, control cells, and =2n phagocytic cells are isolated by using a cell surface target (e.g., receptor protein) on the surface of the cells. In some embodiments, the cell surface target is a protein that has been engulfed by >2n phagocytic cells. In some embodiments, the cell surface target is expressed by cells on their plasma membranes. In some embodiments, the cell surface target is an exogenous protein that is translocated on the plasma membranes, but not expressed by the cells (e.g., the >2n phagocytic cells). In some embodiments, the cell surface target is a marker of the disease or condition to be detected.

[0075] In certain embodiments, the circulating vesicles are isolated using chromatographic isolation, affinity isolation, or ultracentrifugation.

[0076] In certain aspects of the methods described herein, analytes include nucleic acids, proteins, lipids, carbohydrates, metabolites, or any combinations of these. In certain aspects of the methods described herein, markers include nucleic acids, proteins, lipids, carbohydrates, metabolites, or any combinations of these. As used herein, the term "nucleic acid" is intended to include DNA molecules (e.g., cDNA or genomic DNA), RNA molecules (e.g., mRNA), DNA-RNA hybrids, and analogs of the DNA or RNA generated using nucleotide analogs. The nucleic acid molecule can be a nucleotide, oligonucleotide, double-stranded DNA, single-stranded DNA, multi-stranded DNA, complementary DNA, genomic DNA, non-coding DNA, messenger RNA (mRNAs), microRNA (miRNAs), small nucleolar RNA (snoRNAs), ribosomal RNA (rRNA), transfer RNA (tRNA), small interfering RNA (siRNA), heterogeneous nuclear RNAs (hnRNA), or small hairpin RNA (shRNA). In some embodiments, the nucleic acid is a transrenal nucleic acid. A transrenal nucleic acid is an extracellular nucleic acid that is excreted in the urine, See, e.g., U.S. Patent Publication No. 20100068711 and U.S. Patent Publication No. 20120021404.

[0077] As used herein, the term "amino acid" includes organic compounds containing both a basic amino group and an acidic carboxyl group. Included within this term are natural amino acids (e.g., L-amino acids), modified and unusual amino acids (e.g., D-amino acids and β-amino acids), as well as amino acids which are known to occur biologically in free or combined form but usually do not occur in proteins. Natural protein occurring amino acids include alanine, arginine, asparagine, aspartic acid, cysteine, glutamic acid, glutamine, glycine, histidine, isoleucine, leucine, lysine, methionine, phenylalanine, serine, threonine, tyrosine, tryptophan, proline, and valine. Natural non-protein amino acids include arginosuccinic acid, citrulline, cysteine sulfuric acid, 3,4-dihydroxyphenylalanine, homocysteine, homoserine, ornithine, 3-monoiodotyrosine, 3,5-diiodotryosine, 3, 5, 5-triiodothyronine, and 3,3',5,5'- tetraiodothyronine. Modified or unusual amino acids include D-amino acids, hydroxylysine, 4-hydroxyproline, N-Cbz-protected amino acids, 2,4-diaminobutyric acid, homoarginine, norleucine, N-methylaminobutyric

acid, naphthylalanine, phenylglycine, .alpha.-phenylproline, tert-leucine, 4-aminocyclohexylalanine, N-methyl-norleucine, 3 ,4-dehydroproline, N,N-dimethylaminoglycine, N-methylaminoglycine, 4-aminopiperidine-4-carboxylic acid, 6-aminocaproic acid, trans-4-(aminomethyl)-cyclohexanecarboxylic acid, 2-, 3-, and 4-(aminomethyl)- benzoic acid, 1-aminocyclopentanecarboxylic acid, 1-aminocyclopropanecarboxylic acid, and 2-benzyl-5-aminopentanoic acid.

[0078] As used herein, the term "peptide" includes compounds that consist of two or more amino acids that are linked by means of a peptide bond. Peptides may have a molecular weight of less than 10,000 Daltons, less than 5,000 Daltons, or less than 2,500 Daltons. The term "peptide" also includes compounds containing both peptide and non-peptide components, such as pseudopeptide or peptidomimetic residues or other non-amino acid components. Such compounds containing both peptide and non-peptide components may also be referred to as a "peptide analog."

[0079] As used herein, the term "protein" includes compounds that consist of amino acids arranged in a linear chain and joined together by peptide bonds between the carboxyl and amino groups of adjacent amino acid residues. Proteins used in methods of the invention include, but are not limited to, amino acids, peptides, antibodies, antibody fragments, cytokines, lipoproteins, or glycoproteins.

[0080] As used herein, the term "antibody" includes polyclonal antibodies, monoclonal antibodies (including full length antibodies which have an immunoglobulin Fc region), antibody compositions with polyepitopic specificity, multispecific antibodies (e.g., bispecific antibodies, diabodies, and single-chain molecules, and antibody fragments (*e.g.*, Fab or $F(ab')_2$, and Fv). For the structure and properties of the different classes of antibodies, see e.g., Basic and Clinical Immunology, 8th Edition, Daniel P. Sties, Abba I. Terr and Tristram G. Parsolw (eds), Appleton & Lange, Norwalk, Conn., 1994, page 71 and Chapter 6.

[0081] As used herein, the term "cytokine" refers to a secreted protein or active fragment or mutant thereof that modulates the activity of cells of the immune system. Examples of cytokines include, without limitation, interleukins, interferons,

chemokines, tumor necrosis factors, colony-stimulating factors for immune cell precursors, and the like.

[0082] As used herein, the term "lipoprotein" includes negatively charged compositions that comprise a core of hydrophobic cholesteryl esters and triglyceride surrounded by a surface layer of amphipathic phospholipids with which free cholesterol and apolipoproteins are associated. Lipoproteins may be characterized by their density (e.g. very-low-density lipoprotein (VLDL), low-density lipoprotein (LDL) and high density lipoprotein (HDL)), which is determined by their size, the relative amounts of lipid and protein. Lipoproteins may also be characterized by the presence or absence of particular modifications (e.g. oxidization, acetylation, or glycation).

[0083] As used herein, the term "glycoprotein" includes glycosides which have one or more oligo- or polysaccharides covalently attached to a peptide or protein. Exemplary glycoproteins can include, without limitation, immunoglobulins, members of the major histocompatibility complex, collagens, mucins, glycoprotein IIb/IIIa, glycoprotein-41 (gp41) and glycoprotein-120 (gp12), follicle-stimulating hormone, alpha-fetoprotein, erythropoietin, transferrins, alkaline phosphatase, and lectins.

[0084] As used herein, the term "lipid" includes synthetic or naturally-occurring compounds which are generally amphipathic and biocompatible. Lipids typically comprise a hydrophilic component and a hydrophobic component. Exemplary lipids include, but are not limited to fatty acids, neutral fats, phosphatides, cholesterol, cholesterol esters, triglycerides, glycolipids, glycerolipids, glycerophospholipids, sphingolipids, sterol lipids, prenol lipids, saccharolipids, polyketides, choline glycerophospholipid, ethanolamine glycerophospholipid, phosphatidylinositol, phosphatidylglycerol, phosphatidylserine, lyso-choline glycerophospholipid, lyso-ethanolamine glycerophospholipid, phosphatidic acid, lyso-phosphatidic acid, sphingomyelin, galactosylceramide, glucosylceramide, sulfatide, free fatty acids, prostaglandins, triacylglycerol, diacylglycerol, monoacylglycerol, acyl-CoA, acylcarnitine, oxysterol, ceramide, cardiolipin, sphingoid base-1-phosphate, shingosine, lyso-sphingomyelin,, gangliosides, plasmalogen, sulfatide, ceramide, low

density lipoproteins (LDLs), very low density lipoproteins (VLDLs), high density lipoproteins (HDLs), sphingoid base-1-phosphates or derivatives thereof.

[0085] As used herein, the term "carbohydrate" includes, but is not limited to, compounds that contain oxygen, hydrogen and carbon atoms, typically $(CH_2O)_n$ wherein n is an integer. Exemplary carbohydrates include, but are not limited to, monosaccharides, disaccharides, polysaccharides, or oligosaccharides.

[0086] As used herein, the term "metabolite" includes any molecule used in metabolism. Metabolites can be products, substrates, or intermediates in metabolic processes. Included within this term are primary metabolites, secondary metabolites, organic metabolites, or inorganic metabolites. Metabolites include, without limitation, amino acids, peptides, acylcarnitines, monosaccharides, lipids and phospholipids, prostaglandins, hydroxyeicosatetraenoic acids, hydroxyoctadecadienoic acids, steroids, bile acids, and glycolipids and phospholipids. Exemplary metabolites can be sphingolipids, glycosphingolipids, sphingosine, ceramide, sphingomyelin, sphingosylphosphorylcholin, dihydrosphingosine, phoshatidylcholine, phosphatidylinositol, phosphatidylserine, lysophoshatidylcholine, lysophosphatidylinositol, lysophosphatidylserine, plasmenylphoshatidylcholine, plasmanylphoshatidylcholine, proteinogenic amino acids, Alanine, Aspartic acid, Glutamic acid, Phenylalanine, Glycine, Histidine, Leucine, Isoleucine, Lysine, Methionine, Proline, Arginine, Serine, Threonine, Valine, Tryptophan, Tyrosine, asymmetrical dimethyl arginine, symmetrical dimethyl arginine, Glutamine, Asparagine, Nitrotyrosine, Hydroxyproline, Kynurenine, 3-Hydroxy kynurenine, non-proteinogenic amino acids, Ornithine, Citrulline, acylcarnitines, carnitine, free carnitine, acylcarnitine, hydroxylacylcarnitine, dicarboxylacylcarnitines, reducing monosaccharides, hexose, pentose, deoxyhexose, creatinine, creatine, spermidine spermine, putrescine, dopamine, serotonin, prostaglandins, hydoxyeicosatetraeneoic acid, Hydroxyoctadecadienoic acid, leukatrienes, thromboxanes, bile acids, sterols, cholesterols, vitamins and cofactors, drugs, and drug metabolites.

[0087] In some embodiments of the invention, a sample may comprise one or more stabilizers for a cell or an analyte such as DNA, RNA, protein, and/or lipid. For

example, a sample may comprise a DNA stabilizer, an RNA stabilizer, and/or a protein stabilizer. Stabilizers are well known in the art and include, for example, DNAse inhibitors, RNAse inhibitors, and protease inhibitors or equivalents thereof.

[0088] In some embodiments of the invention, profiles of at least one or more markers of a disease or condition are compared. This comparison can be quantitative or qualitative. Quantitative measurements can be taken using any of the assays described herein. For example, sequencing, direct sequencing, random shotgun sequencing, Sanger dideoxy termination sequencing, targeted sequencing, whole-genome sequencing, sequencing by hybridization, pyrosequencing, capillary electrophoresis, gel electrophoresis, duplex sequencing, cycle sequencing, single-base extension sequencing, solid-phase sequencing, high-throughput sequencing, massively parallel signature sequencing, emulsion PCR, co-amplification at lower denaturation temperature-PCR (COLD-PCR), sequencing by reversible dye terminator, paired-end sequencing, near-term sequencing, exonuclease sequencing, sequencing by ligation, short-read sequencing, single-molecule sequencing, sequencing-by-synthesis, real-time sequencing, reverse-terminator sequencing, nanopore sequencing, 454 sequencing, Solexa Genome Analyzer sequencing, SOLiD® sequencing, MS-PET sequencing, mass spectrometry, matrix assisted laser desorption/ionization-time of flight (MALDI-TOF) mass spectrometry, electrospray ionization (ESI) mass spectrometry, surface-enhanced laser deorption/ionization-time of flight (SELDI-TOF) mass spectrometry, quadrupole-time of flight (Q-TOF) mass spectrometry, atmospheric pressure photoionization mass spectrometry (APPI-MS), Fourier transform mass spectrometry (FTMS), matrix-assisted laser desorption/ionization-Fourier transform-ion cyclotron resonance (MALDI-FT-ICR) mass spectrometry, secondary ion mass spectrometry (SIMS), polymerase chain reaction (PCR) analysis, quantitative PCR, real-time PCR, fluorescence assay, colorimetric assay, chemiluminescent assay, or a combination thereof.

[0089] Quantitative comparisons can include statistical analyses such as t-test, ANOVA, Krustal-Wallis, Wilcoxon, Mann-Whitney, and odds ratio. Quantitative differences can include differences in the levels of markers between profiles or differences in the numbers of markers present between profiles, and combinations

thereof. Examples of levels of the markers can be, without limitation, gene expression levels, nucleic acid levels, protein levels, lipid levels, and the like. Qualitative differences can include, but are not limited to, activation and inactivation, protein degradation, nucleic acid degradation, and covalent modifications.

[0090] In certain embodiments of the invention, the profile is a nucleic acid profile, a protein profile, a lipid profile, a carbohydrate profile, a metabolite profile, or a combination thereof. The profile can be qualitatively or quantitatively determined.

[0091] A nucleic acid profile can be, without limitation, a genotypic profile, a single nucleotide polymorphism profile, a gene mutation profile, a gene copy number profile, a DNA methylation profile, a DNA acetylation profile, a chromosome dosage profile, a gene expression profile, or a combination thereof.

[0092] The nucleic acid profile can be determined by any methods known in the art to detect genotypes, single nucleotide polymorphisms, gene mutations, gene copy numbers, DNA methylation states, DNA acetylation states, chromosome dosages. Exemplary methods include, but are not limited to, polymerase chain reaction (PCR) analysis, sequencing analysis, electrophoretic analysis, restriction fragment length polymorphism (RFLP) analysis, Northern blot analysis, quantitative PCR, reverse-transcriptase-PCR analysis (RT-PCR), allele-specific oligonucleotide hybridization analysis, comparative genomic hybridization, heteroduplex mobility assay (HMA), single strand conformational polymorphism (SSCP), denaturing gradient gel electrophisis (DGGE), RNAase mismatch analysis, mass spectrometry, tandem mass spectrometry, matrix assisted laser desorption/ionization-time of flight (MALDI-TOF) mass spectrometry, electrospray ionization (ESI) mass spectrometry, surface-enhanced laser deorption/ionization-time of flight (SELDI-TOF) mass spectrometry, quadrupole-time of flight (Q-TOF) mass spectrometry, atmospheric pressure photoionization mass spectrometry (APPI-MS), Fourier transform mass spectrometry (FTMS), matrix-assisted laser desorption/ionization-Fourier transform-ion cyclotron resonance (MALDI-FT-ICR) mass spectrometry, secondary ion mass spectrometry (SIMS), surface plasmon resonance, Southern blot analysis, in situ hybridization, fluorescence in situ hybridization (FISH), chromogenic in situ hybridization (CISH),

immunohistochemistry (IHC), microarray, comparative genomic hybridization, karyotyping, multiplex ligation-dependent probe amplification (MLPA), Quantitative Multiplex PCR of Short Fluorescent Fragments (QMPSF), microscopy, methylation specific PCR (MSP) assay, HpaII tiny fragment Enrichment by Ligation-mediated PCR (HELP) assay, radioactive acetate labeling assays, colorimetric DNA acetylation assay, chromatin immunoprecipitation combined with microarray (ChIP-on-chip) assay, restriction landmark genomic scanning, Methylated DNA immunoprecipitation (MeDIP), molecular break light assay for DNA adenine methyltransferase activity, chromatographic separation, methylation-sensitive restriction enzyme analysis, bisulfite-driven conversion of non-methylated cytosine to uracil, co-amplification at lower denaturation temperature-PCR (COLD-PCR), multiplex PCR, methyl-binding PCR analysis, or a combination thereof.

[0093] As used herein, the term "sequencing" is used in a broad sense and refers to any technique known in the art that allows the order of at least some consecutive nucleotides in at least part of a nucleic acid to be identified, including without limitation at least part of an extension product or a vector insert. Exemplary sequencing techniques include targeted sequencing, single molecule real-time sequencing, electron microscopy-based sequencing, transistor-mediated sequencing, direct sequencing, random shotgun sequencing, Sanger dideoxy termination sequencing, exon sequencing, whole-genome sequencing, sequencing by hybridization, pyrosequencing, capillary electrophoresis, gel electrophoresis, duplex sequencing, cycle sequencing, single-base extension sequencing, solid-phase sequencing, high-throughput sequencing, massively parallel signature sequencing, emulsion PCR, co-amplification at lower denaturation temperature-PCR (COLD-PCR), multiplex PCR, sequencing by reversible dye terminator, paired-end sequencing, near-term sequencing, exonuclease sequencing, sequencing by ligation, short-read sequencing, single-molecule sequencing, sequencing-by-synthesis, real-time sequencing, reverse-terminator sequencing, nanopore sequencing, 454 sequencing, Solexa Genome Analyzer sequencing, SOLiD® sequencing, MS-PET sequencing, mass spectrometry, and a combination thereof.. In some embodiments, sequencing comprises an detecting the sequencing product using an instrument, for example but not limited to an ABI PRISM® 377 DNA Sequencer, an ABI PRISM®

310, 3100, 3100-Avant, 3730, or 373OxI Genetic Analyzer, an ABI PRISM® 3700 DNA Analyzer, or an Applied Biosystems SOLiD™ System (all from Applied Biosystems), a Genome Sequencer 20 System (Roche Applied Science), or a mass spectrometer. In certain embodiments, sequencing comprises emulsion PCR. In certain embodiments, sequencing comprises a high throughput sequencing technique, for example but not limited to, massively parallel signature sequencing (MPSS).

[0094] In further embodiments of the invention, a protein profile can be a protein expression profile, a protein activation profile, or a combination thereof. In some embodiments, a protein activation profile can comprise determining a phosphorylation state, an ubiquitination state, a myristoylation state, or a conformational state of the protein.

[0095] A protein profile can be detected by any methods known in the art for detecting protein expression levels, protein phosphorylation state, protein ubiquitination state, protein myristoylation state, or protein conformational state. In some embodiments, a protein profile can be determined by an immunohistochemistry assay, an enzyme-linked immunosorbent assay (ELISA), in situ hybridization, chromatography, liquid chromatography, size exclusion chromatography, high performance liquid chromatography (HPLC), gas chromatography, mass spectrometry, tandem mass spectrometry, matrix assisted laser desorption/ionization-time of flight (MALDI-TOF) mass spectrometry, electrospray ionization (ESI) mass spectrometry, surface-enhanced laser deorption/ionization-time of flight (SELDI-TOF) mass spectrometry, quadrupole-time of flight (Q-TOF) mass spectrometry, atmospheric pressure photoionization mass spectrometry (APPI-MS), Fourier transform mass spectrometry (FTMS), matrix-assisted laser desorption/ionization-Fourier transform-ion cyclotron resonance (MALDI-FT-ICR) mass spectrometry, secondary ion mass spectrometry (SIMS), radioimmunoassays, microscopy, microfluidic chip-based assays, surface plasmon resonance, sequencing, Western blotting assay, or a combination thereof.

[0096] In some embodiments of the invention, a lipid profile can be determined by chromatography, liquid chromatography, size exclusion chromatography, high

performance liquid chromatography (HPLC), gas chromatography, mass spectrometry, tandem mass spectrometry, matrix assisted laser desorption/ionization-time of flight (MALDI-TOF) mass spectrometry, electrospray ionization (ESI) mass spectrometry, surface-enhanced laser deorption/ionization-time of flight (SELDI-TOF) mass spectrometry, quadrupole-time of flight (Q-TOF) mass spectrometry, atmospheric pressure photoionization mass spectrometry (APPI-MS), Fourier transform mass spectrometry (FTMS), matrix-assisted laser desorption/ionization-Fourier transform-ion cyclotron resonance (MALDI-FT-ICR) mass spectrometry, secondary ion mass spectrometry (SIMS), radioimmunoassays, microfluidic chip-based assay, detection of fluorescence, detection of chemiluminescence, or a combination thereof. Further methods for analyzing lipid content in a biological sample are known in the art (See, e.g., Kang et al. (1992) Biochim. Biophys. Acta. 1128:267; Weylandt et al. (1996) Lipids 31 :977; J. Schiller et al. (1999) Anal. Biochem. 267:46; Kang et al. (2001) Proc. Natl. Acad. Sci. USA 98:4050; Schiller et al. (2004) Prog. Lipid Res. 43:499). One exemplary method of lipid analysis is to extract lipids from a biological sample (e.g. using chloroform-methanol (2:1, vol/vol) containing 0.005% butylated hydroxytoluene (BHT, as an antioxidant)), prepare fatty acid methyl esters (e.g., using 14% BF3-methanol reagent), and quantify the fatty acid methyl esters (e.g., by HPLC, TLC, by gas chromatography-mass spectroscopy using commercially available gas chromatographs, mass spectrometers, and/or combination gas chromatograph/mass spectrometers). Fatty acid mass is determined by comparing areas of various analyzed fatty acids to that of a fixed concentration of internal standard.

[0097] In some embodiments of the invention, a carbohydrate profile can be determined by chromatography, liquid chromatography, size exclusion chromatography, high performance liquid chromatography (HPLC), gas chromatography, mass spectrometry, tandem mass spectrometry, matrix assisted laser desorption/ionization-time of flight (MALDI-TOF) mass spectrometry, electrospray ionization (ESI) mass spectrometry, surface-enhanced laser deorption/ionization-time of flight (SELDI-TOF) mass spectrometry, quadrupole-time of flight (Q-TOF) mass spectrometry, atmospheric pressure photoionization mass spectrometry (APPI-MS), Fourier transform mass spectrometry (FTMS), matrix-assisted laser

desorption/ionization-Fourier transform-ion cyclotron resonance (MALDI-FT-ICR) mass spectrometry, secondary ion mass spectrometry (SIMS), radioimmunoassays, microfluidic chip-based assay, detection of fluorescence, detection of chemiluminescence, or a combination thereof.

[0098] In some embodiments of the invention, a metabolite profile can be determind by chromatography, liquid chromatography, size exclusion chromatography, high performance liquid chromatography (HPLC), gas chromatography, mass spectrometry, tandem mass spectrometry, matrix assisted laser desorption/ionization-time of flight (MALDI-TOF) mass spectrometry, electrospray ionization (ESI) mass spectrometry, surface-enhanced laser deorption/ionization-time of flight (SELDI-TOF) mass spectrometry, quadrupole-time of flight (Q-TOF) mass spectrometry, atmospheric pressure photoionization mass spectrometry (APPI-MS), Fourier transform mass spectrometry (FTMS), matrix-assisted laser desorption/ionization-Fourier transform-ion cyclotron resonance (MALDI-FT-ICR) mass spectrometry, secondary ion mass spectrometry (SIMS), radioimmunoassays, microfluidic chip-based assay, detection of fluorescence, detection of chemiluminescence, or a combination thereof.

[0099] As used herein, the "difference" between different profiles detected by the methods of this invention can refer to different gene copy numbers, different DNA, RNA, protein, lipid, or carbohydrate expression levels, different DNA methylation states, different DNA acetylation states, and different protein modification states. The difference can be a difference greater than 1 fold. In some embodiments, the difference is a 1.05-fold, 1.1-fold, 1.2-fold, 1.3-fold, 1.4-fold, 1.5-fold, 2-fold, 2.5-fold, 3-fold, 4-fold, 5-fold, 6-fold, 7-fold, 8-fold, 9-fold, or 10-fold difference. In some embodiments, the difference is any fold difference between 1-10, 2-10, 5-10, 10-20, or 10-100 fold.

[0100] In some embodiments, the difference is differential gene expression (DGE), e.g. DGE of phagocytes vs. non-phagocytes or >2n phagocytes vs. =2n phagocytes. DGE can be measured as $X = \log_2(Y_P) - \log_2(Y_{NP})$. The DGE may be any number, provided that it is significantly different between the combination sample and the =2n

phagocytes, non-phagocytes, control cells, or repository of markers. For example, a 2-fold increased in gene expression could be represented as $X = \log_2(Y_P) - \log_2(Y_{NP}) = \log_2(Y_P/Y_{NP}) = \log_2(2) = 1$, while a 2-fold decrease in gene expression could be represented as $X = \log_2(Y_P) - \log_2(Y_{NP}) = \log_2(Y_P/Y_{NP}) = \log_2(1/2) = -1$. Down-regulated genes have $X < 0$, while up-regulated genes have $X > 0$. See, *e.g.,* Efron, J Am Stat Assoc 104:1015-1028 (2009).

[0101] A general principle of assays to detect markers involves preparing a sample or reaction mixture that may contain the marker (e.g., one or more of DNA, RNA, protein, polypeptide, carbohydrate, lipid, metabolite, and the like) and a probe under appropriate conditions and for a time sufficient to allow the marker and probe to interact and bind, thus forming a complex that can be removed and/or detected in the reaction mixture. These assays can be conducted in a variety of ways.

[0102] For example, one method to conduct such an assay would involve anchoring the marker or probe onto a solid phase support, also referred to as a substrate, and detecting target marker/probe complexes anchored on the solid phase at the end of the reaction. In one embodiment of such a method, a sample from a subject, which is to be assayed for presence and/or concentration of marker, can be anchored onto a carrier or solid phase support. In another embodiment, the reverse situation is possible, in which the probe can be anchored to a solid phase and a sample from a subject can be allowed to react as an unanchored component of the assay.

[0103] There are many established methods for anchoring assay components to a solid phase. These include, without limitation, marker or probe molecules which are immobilized through conjugation of biotin and streptavidin. Such biotinylated assay components can be prepared from biotin-NHS(N-hydroxy-succinimide) using techniques known in the art (e.g., biotinylation kit, Pierce Chemicals, Rockford, IL), and immobilized in the wells of streptavidin-coated 96 well plates (Pierce Chemical). In certain embodiments, the surfaces with immobilized assay components can be prepared in advance and stored.

[0104] Other suitable carriers or solid phase supports for such assays include any material capable of binding the class of molecule to which the marker or probe

belongs. Well known supports or carriers include, but are not limited to, glass, polystyrene, nylon, polypropylene, nylon, polyethylene, dextran, amylases, natural and modified celluloses, polyacrylamides, gabbros, and magnetite.

[0105] In order to conduct assays with the above mentioned approaches, the non-immobilized component is added to the solid phase upon which the second component is anchored. After the reaction is complete, uncomplexed components may be removed (e.g., by washing) under conditions such that any complexes formed will remain immobilized upon the solid phase. The detection of marker/probe complexes anchored to the solid phase can be accomplished in a number of methods outlined herein.

[0106] In certain exemplary embodiments, the probe, when it is the unanchored assay component, can be labeled for the purpose of detection and readout of the assay, either directly or indirectly, with detectable labels discussed herein and which are well-known to one skilled in the art.

[0107] It is also possible to directly detect marker/probe complex formation without further manipulation or labeling of either component (marker or probe), for example by utilizing the technique of fluorescence energy transfer (see, for example, U.S. Patent Nos. 5,631,169 and 4,868,103). A fluorophore label on the first, 'donor' molecule is selected such that, upon excitation with incident light of appropriate wavelength, its emitted fluorescent energy will be absorbed by a fluorescent label on a second 'acceptor' molecule, which in turn is able to fluoresce due to the absorbed energy. Alternately, the 'donor' protein molecule may simply utilize the natural fluorescent energy of tryptophan residues. Labels are chosen that emit different wavelengths of light, such that the 'acceptor' molecule label may be differentiated from that of the 'donor'. Since the efficiency of energy transfer between the labels is related to the distance separating the molecules, spatial relationships between the molecules can be assessed. In a situation in which binding occurs between the molecules, the fluorescent emission of the 'acceptor' molecule label in the assay should be maximal. An FET binding event can be conveniently measured through

standard fluorometric detection means well known in the art (e.g., using a fluorimeter).

[0108] In another embodiment, determination of the ability of a probe to recognize a marker can be accomplished without labeling either assay component (probe or marker) by utilizing a technology such as real-time Biomolecular Interaction Analysis (BIA) (see, e.g., Sjolander, S. and Urbaniczky, C, 1991, Anal. Chem. 63:2338 2345 and Szabo et al, 1995, Curr. Opin. Struct. Biol. 5:699 705). As used herein, "BIA" or "surface plasmon resonance" is a technology for studying biospecific interactions in real time, without labeling any of the interactants (e.g., BIAcore). Changes in the mass at the binding surface (indicative of a binding event) result in alterations of the refractive index of light near the surface (the optical phenomenon of surface plasmon resonance (SPR)), resulting in a detectable signal which can be used as an indication of real-time reactions between biological molecules.

[0109] Alternatively, in another embodiment, analogous diagnostic and prognostic assays can be conducted with marker and probe as solutes in a liquid phase. In such an assay, the complexed marker and probe are separated from uncomplexed components by any of a number of standard techniques, including but not limited to: differential centrifugation, chromatography, electrophoresis and immunoprecipitation. In differential centrifugation, marker/probe complexes may be separated from uncomplexed assay components through a series of centrifugal steps, due to the different sedimentation equilibria of complexes based on their different sizes and densities (see, for example, Rivas and Minton (1993) Trends Biochem. Sci. 18:284). Standard chromatographic techniques may also be utilized to separate complexed molecules from uncomplexed ones. For example, gel filtration chromatography separates molecules based on size, and through the utilization of an appropriate gel filtration resin in a column format, for example, the relatively larger complex may be separated from the relatively smaller uncomplexed components. Similarly, the relatively different charge properties of the marker/probe complex as compared to the uncomplexed components may be exploited to differentiate the complex from uncomplexed components, for example through the utilization of ion-exchange chromatography resins. Such resins and chromatographic techniques are well known

to one skilled in the art (see, e.g., Heegaard (1998) J. Mol. Recognit. 11 :141; Hage and Tweed (1997) J. Chromatogr. B. Biomed. Sci. Appl. 12:499). Gel electrophoresis may also be employed to separate complexed assay components from unbound components (see, e.g., Ausubel et al, ed., Current Protocols in Molecular Biology, John Wiley & Sons, New York, 1987 1999). In this technique, protein or nucleic acid complexes are separated based on size or charge, for example. In order to maintain the binding interaction during the electrophoretic process, non-denaturing gel matrix materials and conditions in the absence of reducing agent are typically preferred. Appropriate conditions to the particular assay and components thereof will be well known to one skilled in the art.

[0110] In certain exemplary embodiments, the level of mRNA corresponding to the marker can be determined either by *in situ* and/or by *in vitro* formats in a biological sample using methods known in the art. Many expression detection methods use isolated RNA. For in vitro methods, any RNA isolation technique that does not select against the isolation of mRNA can be utilized for the purification of RNA from blood cells (see, e.g., Ausubel et al, ed., Current Protocols in Molecular Biology, John Wiley & Sons, New York 1987 1999). Additionally, large numbers of cells and/or samples can readily be processed using techniques well known to those of skill in the art, such as, for example, the single-step RNA isolation process of Chomczynski (1989, U.S. Patent No. 4,843,155).

[0111] Isolated mRNA can be used in hybridization or amplification assays that include, but are not limited to, Southern or Northern analyses, polymerase chain reaction analyses and probe arrays. In certain exemplary embodiments, a diagnostic method for the detection of mRNA levels involves contacting the isolated mRNA with a nucleic acid molecule (probe) that can hybridize to the mRNA encoded by the gene being detected. The nucleic acid probe can be, for example, a full-length cDNA, or a portion thereof, such as an oligonucleotide of at least 7, 15, 30, 50, 100, 250 or 500 nucleotides in length and sufficient to specifically hybridize under stringent conditions to an mRNA or genomic DNA encoding a marker of the present invention. Other suitable probes for use in the diagnostic assays of the invention are described

herein. Hybridization of an mRNA with the probe indicates that the marker in question is being expressed.

[0112] In one format, the mRNA is immobilized on a solid surface and contacted with a probe, for example by running the isolated mRNA on an agarose gel and transferring the mRNA from the gel to a membrane, such as nitrocellulose. In an alternative format, the probe(s) are immobilized on a solid surface and the mRNA is contacted with the probe(s), for example, in a gene chip array. A skilled artisan can readily adapt known mRNA detection methods for use in detecting the level of mRNA encoded by the markers of the present invention.

[0113] An alternative method for determining the level of mRNA corresponding to a marker of the present invention in a sample involves the process of nucleic acid amplification, e.g., by RT-PCR (the experimental embodiment set forth in U.S. Patent Nos. 4,683,195 and 4,683,202), COLD-PCR (Li et al. (2008) Nat. Med. 14:579), ligase chain reaction (Barany, 1991, Proc. Natl. Acad. Sci. USA, 88:189), self sustained sequence replication (Guatelli et al., 1990, Proc. Natl. Acad. Sci. USA 87:1874), transcriptional amplification system (Kwoh et al. (1989) Proc. Natl. Acad. Sci. USA 86:1173), Q- Beta Replicase (Lizardi et al. (1988) Bio/Technology 6:1197), rolling circle replication (U.S. Patent No. 5,854,033) or any other nucleic acid amplification method, followed by the detection of the amplified molecules using techniques well known to those of skill in the art. These detection schemes are especially useful for the detection of nucleic acid molecules if such molecules are present in very low numbers. As used herein, amplification primers are defined as being a pair of nucleic acid molecules that can anneal to 5' or 3' regions of a gene (plus and minus strands, respectively, or vice-versa) and contain a short region in between. In general, amplification primers are from about 10 to 30 nucleotides in length and flank a region from about 50 to 200 nucleotides in length. Under appropriate conditions and with appropriate reagents, such primers permit the amplification of a nucleic acid molecule comprising the nucleotide sequence flanked by the primers.

[0114] For *in situ* methods, mRNA does not need to be isolated from the sample (*e.g.*, a bodily fluid (*e.g.*, blood cells)) prior to detection. In such methods, a cell or tissue sample is prepared/processed using known histological methods. The sample is then immobilized on a support, typically a glass slide, and then contacted with a probe that can hybridize to mRNA that encodes the marker.

[0115] As an alternative to making determinations based on the absolute expression level of the marker, determinations may be based on the normalized expression level of the marker. Expression levels are normalized by correcting the absolute expression level of a marker by comparing its expression to the expression of a gene that is not a marker, *e.g.*, a housekeeping gene that is constitutively expressed. Suitable genes for normalization include housekeeping genes such as the actin gene, or epithelial cell-specific genes. This normalization allows the comparison of the expression level in a patient sample from one source to a patient sample from another source, e.g., to compare a combination sample from an individual to a =2n phagocytic or a non-phagocytic blood cell from the individual.

[0116] In one embodiment of this invention, a protein or polypeptide corresponding to a marker is detected. In certain embodiments, an agent for detecting a protein or polypeptide can be an antibody capable of binding to the polypeptide, such as an antibody with a detectable label. As used herein, the term "labeled," with regard to a probe or antibody, is intended to encompass direct labeling of the probe or antibody by coupling (i.e., physically linking) a detectable substance to the probe or antibody, as well as indirect labeling of the probe or antibody by reactivity with another reagent that is directly labeled. Examples of indirect labeling include detection of a primary antibody using a fluorescently labeled secondary antibody and end-labeling of a DNA probe with biotin such that it can be detected with fluorescently labeled streptavidin. Antibodies can be polyclonal or monoclonal. An intact antibody, or a fragment thereof (*e.g.*, Fab or F(ab')2) can be used. In one format, antibodies, or antibody fragments, can be used in methods such as Western blots or immunofluorescence techniques to detect the expressed proteins. In such uses, it is generally preferable to immobilize either the antibody or proteins on a solid support. Suitable solid phase supports or carriers include any support capable of binding an antigen or an antibody.

Well known supports or carriers include glass, polystyrene, polypropylene, polyethylene, dextran, nylon, amylases, natural and modified celluloses, polyacrylamides, gabbros, magnetite and the like.

[0117] A variety of formats can be employed to determine whether a sample contains a protein that binds to a given antibody. Examples of such formats include, but are not limited to, competitive and non-competitive immunoassay, enzyme immunoassay (EIA), radioimmunoassay (RIA), antigen capture assays, two-antibody sandwich assays, Western blot analysis, enzyme linked immunoabsorbant assay (ELISA), a planar array, a colorimetric assay, a chemiluminescent assay, a fluorescent assay, and the like. Immunoassays, including radioimmmunoassays and enzyme-linked immunoassays, are useful in the methods of the present invention. A skilled artisan can readily adapt known protein/antibody detection methods for use in determining whether cells (e.g., bodily fluid cells such as blood cells) express a marker of the present invention.

[0118] One skilled in the art will know many other suitable carriers for binding antibody or antigen, and will be able to adapt such support for use with the present invention. For example, protein isolated from cells (e.g., bodily fluid cells such as blood cells) can be run on a polyacrylamide gel electrophoresis and immobilized onto a solid phase support such as nitrocellulose. The support can then be washed with suitable buffers followed by treatment with the detectably labeled antibody. The solid phase support can then be washed with the buffer a second time to remove unbound antibody. The amount of bound label on the solid support can then be detected by conventional means.

[0119] In certain exemplary embodiments, assays are provided for diagnosis, prognosis, assessing the risk of developing a disease, assessing the efficacy of a treatment, monitoring the progression or regression of a disease, and identifying a compound capable of ameliorating or treating a disease. An exemplary method for these methods involves obtaining a bodily fluid sample from a test subject and contacting the bodily fluid sample with a compound or an agent capable of detecting one or more of the markers of the disease or condition, *e.g.*, marker nucleic acid (*e.g.*,

mRNA, genomic DNA), marker peptide (*e.g.*, polypeptide or protein), marker lipid (*e.g.*, cholesterol), or marker metabolite (*e.g.*, creatinine) such that the presence of the marker is detected in the biological sample. In one embodiment, an agent for detecting marker mRNA or genomic DNA is a labeled nucleic acid probe capable of hybridizing to marker mRNA or genomic DNA. The nucleic acid probe can be, for example, a full-length marker nucleic acid or a portion thereof. Other suitable probes for use in the diagnostic assays of the invention are described herein.

[0120] As used herein, a compound capable of ameliorating or treating a disease or condition can include, without limitations, any substance that can improve symptoms or prognosis, prevent progression of the disease or condition, promote regression of the disease or condition, or eliminate the disease or condition.

[0121] The methods of the invention can also be used to detect genetic alterations in a marker gene, thereby determining if a subject with the altered gene is at risk for developing a disease and/or disorder associated with cancer and/or an infectious agent, and/or one or more other disorders described herein characterized by misregulation in a marker protein activity or nucleic acid expression, such as cancer. In certain embodiments, the methods include detecting, in a cell free bodily fluid sample from the subject, the presence or absence of a genetic alteration characterized by an alteration affecting the integrity of a gene encoding a marker peptide and/or a marker gene. For example, such genetic alterations can be detected by ascertaining the existence of at least one of: 1) a deletion of one or more nucleotides from one or more marker genes; 2) an addition of one or more nucleotides to one or more marker genes; 3) a substitution of one or more nucleotides of one or more marker genes, 4) a chromosomal rearrangement of one or more marker genes; 5) an alteration in the level of a messenger RNA transcript of one or more marker genes; 6) aberrant modification of one or more marker genes, such as of the methylation pattern of the genomic DNA; 7) the presence of a non-wild type splicing pattern of a messenger RNA transcript of one or more marker genes; 8) a non-wild type level of a one or more marker proteins; 9) allelic loss of one or more marker genes; and 10) inappropriate post-translational modification of one or more marker proteins. As described herein, there are a large

number of assays known in the art which can be used for detecting alterations in one or more marker genes.

[0122] In certain embodiments, detection of the alteration involves the use of a probe/primer in a polymerase chain reaction (PCR) (see, e.g., U.S. Patent Nos. 4,683,195, 4,683,202 and 5,854,033), such as real-time PCR, COLD-PCR (Li et al. (2008) Nat. Med. 14:579), anchor PCR, recursive PCR or RACE PCR, or, alternatively, in a ligation chain reaction (LCR) (see, e.g., Landegran et al. (1988) Science 241:1077; Prodromou and Pearl (1992) Protein Eng. 5:827; and Nakazawa et al. (1994) Proc. Natl. Acad. Sci. USA 91:360), the latter of which can be particularly useful for detecting point mutations in a marker gene (see Abravaya et al. (1995) Nucleic Acids Res. 23:675). This method can include the steps of collecting a sample of cell free bodily fluid from a subject, isolating nucleic acid (e.g., genomic, mRNA or both) from the sample, contacting the nucleic acid sample with one or more primers which specifically hybridize to a marker gene under conditions such that hybridization and amplification of the marker gene (if present) occurs, and detecting the presence or absence of an amplification product, or detecting the size of the amplification product and comparing the length to a control sample. It is anticipated that PCR and/or LCR may be desirable to use as a preliminary amplification step in conjunction with any of the techniques used for detecting mutations described herein.

[0123] Alternative amplification methods include: self sustained sequence replication (Guatelli et al., (1990) Proc. Natl. Acad. Sci. USA 87:1874), transcriptional amplification system (Kwoh et al., (1989) Proc. Natl. Acad. Sci. USA 86:1173), Q Beta Replicase (Lizardi et al. (1988) Bio-Technology 6:1197), or any other nucleic acid amplification method, followed by the detection of the amplified molecules using techniques well known to those of skill in the art. These detection schemes are especially useful for the detection of nucleic acid molecules if such molecules are present in very low numbers.

[0124] In an alternative embodiment, mutations in one or more marker genes from a sample can be identified by alterations in restriction enzyme cleavage patterns. For example, sample and control DNA is isolated, optionally amplified, digested with one

or more restriction endonucleases, and fragment length sizes are determined by gel electrophoresis and compared. Differences in fragment length sizes between sample and control DNA indicates mutations in the sample DNA. Moreover, the use of sequence specific ribozymes (see, for example, U.S. Pat. No. 5,498,531) can be used to score for the presence of specific mutations by development or loss of a ribozyme cleavage site.

[0125] In other embodiments, genetic mutations in one or more of the markers described herein can be identified by hybridizing a sample and control nucleic acids, e.g., DNA or RNA, to high density arrays containing hundreds or thousands of oligonucleotides probes (Cronin et al. (1996) Human Mutation 7: 244; Kozal et al. (1996) Nature Medicine 2:753). For example, genetic mutations in a marker nucleic acid can be identified in two dimensional arrays containing light-generated DNA probes as described in Cronin, M. T. et al. supra. Briefly, a first hybridization array of probes can be used to scan through long stretches of DNA in a sample and control to identify base changes between the sequences by making linear arrays of sequential overlapping probes. This step allows the identification of point mutations. This step is followed by a second hybridization array that allows the characterization of specific mutations by using smaller, specialized probe arrays complementary to all variants or mutations detected. Each mutation array is composed of parallel probe sets, one complementary to the wild-type gene and the other complementary to the mutant gene.

[0126] In yet another embodiment, any of a variety of sequencing reactions known in the art can be used to directly sequence a marker gene and detect mutations by comparing the sequence of the sample marker gene with the corresponding wild-type (control) sequence. Examples of sequencing reactions include those based on techniques developed by Maxam and Gilbert ((1977) Proc. Natl. Acad. Sci. USA 74:560) or Sanger ((1977) Proc. Natl. Acad. Sci. USA 74:5463). It is also contemplated that any of a variety of automated sequencing procedures can be utilized when performing the diagnostic assays ((1995) Biotechniques 19:448), including sequencing by mass spectrometry (see, e.g., PCT International Publication No. WO

94/16101; Cohen et al. (1996) Adv. Chromatogr. 36:127-162; and Griffin et al. (1993) Appl. Biochem. Biotechnol. 38:147).

[0127] Other methods for detecting mutations in a marker gene include methods in which protection from cleavage agents is used to detect mismatched bases in RNA/RNA or RNA/DNA heteroduplexes (Myers et al. (1985) Science 230:1242). In general, the art technique of "mismatch cleavage" starts by providing heteroduplexes formed by hybridizing (labeled) RNA or DNA containing the wild-type marker sequence with potentially mutant RNA or DNA obtained from a tissue sample. The double-stranded duplexes are treated with an agent which cleaves single-stranded regions of the duplex such as which will exist due to base pair mismatches between the control and sample strands. For instance, RNA/DNA duplexes can be treated with RNase and DNA/DNA hybrids treated with S1 nuclease to enzymatically digesting the mismatched regions. In other embodiments, either DNA/DNA or RNA/DNA duplexes can be treated with hydroxylamine or osmium tetroxide and with piperidine in order to digest mismatched regions. After digestion of the mismatched regions, the resulting material is then separated by size on denaturing polyacrylamide gels to determine the site of mutation. See, for example, Cotton et al. (1988) Proc. Natl. Acad. Sci. USA 85:4397; Saleeba et al. (1992) Methods Enzymol. 217:286. In one embodiment, the control DNA or RNA can be labeled for detection.

[0128] In still another embodiment, the mismatch cleavage reaction employs one or more proteins that recognize mismatched base pairs in double-stranded DNA (so called "DNA mismatch repair" enzymes) in defined systems for detecting and mapping point mutations in marker cDNAs obtained from samples of cells. For example, the mutY enzyme of E. coli cleaves A at G/A mismatches and the thymidine DNA glycosylase from HeLa cells cleaves T at G/T mismatches (Hsu et al. (1994) Carcinogenesis 15:1657). According to an exemplary embodiment, a probe based on a marker sequence, e.g., a wild-type marker sequence, is hybridized to a cDNA or other DNA product from a test cell(s). The duplex is treated with a DNA mismatch repair enzyme, and the cleavage products, if any, can be detected from electrophoresis protocols or the like. See, for example, U.S. Patent No. 5,459,039.

[0129] In other embodiments, alterations in electrophoretic mobility will be used to identify mutations in marker genes. For example, single strand conformation polymorphism (SSCP) may be used to detect differences in electrophoretic mobility between mutant and wild type nucleic acids (Orita et al. (1989) Proc. Natl. Acad. Sci. USA 86:2766, see also Cotton (1993) Mutat. Res. 285:125; and Hayashi (1992) Genet. Anal. Tech. Appl. 9:73). Single-stranded DNA fragments of sample and control marker nucleic acids will be denatured and allowed to renature. The secondary structure of single-stranded nucleic acids varies according to sequence, the resulting alteration in electrophoretic mobility enables the detection of even a single base change. The DNA fragments may be labeled or detected with labeled probes. The sensitivity of the assay may be enhanced by using RNA (rather than DNA), in which the secondary structure is more sensitive to a change in sequence. In one embodiment, the subject method utilizes heteroduplex analysis to separate double stranded heteroduplex molecules on the basis of changes in electrophoretic mobility (Keen et al. (1991) Trends Genet. 7:5).

[0130] In yet another embodiment the movement of mutant or wild-type fragments in polyacrylamide gels containing a gradient of denaturant is assayed using denaturing gradient gel electrophoresis (DGGE) (Myers et al. (1985) Nature 313:495). When DGGE is used as the method of analysis, DNA will be modified to insure that it does not completely denature, for example by adding a GC clamp of approximately 40 bp of high-melting GC-rich DNA by PCR. In a further embodiment, a temperature gradient is used in place of a denaturing gradient to identify differences in the mobility of control and sample DNA (Rosenbaum and Reissner (1987) Biophys. Chem. 265:12753).

[0131] Examples of other techniques for detecting point mutations include, but are not limited to, selective oligonucleotide hybridization, selective amplification or selective primer extension. For example, oligonucleotide primers may be prepared in which the known mutation is placed centrally and then hybridized to target DNA under conditions which permit hybridization only if a perfect match is found (Saiki et al. (1986) Nature 324:163; Saiki et al. (1989) Proc. Natl. Acad. Sci. USA 86:6230). Such allele specific oligonucleotides are hybridized to PCR amplified target DNA or a

number of different mutations when the oligonucleotides are attached to the hybridizing membrane and hybridized with labeled target DNA.

[0132] Alternatively, allele specific amplification technology which depends on selective PCR amplification may be used in conjunction with the instant invention. Oligonucleotides used as primers for specific amplification may carry the mutation of interest in the center of the molecule (so that amplification depends on differential hybridization) (Gibbs et al. (1989) Nucl. Acids Res. 17:2437) or at the extreme 3' end of one primer where, under appropriate conditions, mismatch can prevent, or reduce polymerase extension (Prossner (1993) Tibtech 11:238). In addition it may be desirable to introduce a novel restriction site in the region of the mutation to create cleavage-based detection (Gasparini et al. (1992) Mol. Cell Probes 6:1). It is anticipated that in certain embodiments amplification may also be performed using Taq ligase for amplification (Barany (1991) Proc. Natl. Acad. Sci. USA 88:189). In such cases, ligation will occur only if there is a perfect match at the 3' end of the 5' sequence making it possible to detect the presence of a known mutation at a specific site by looking for the presence or absence of amplification.

[0133] In some embodiments, this invention provides a method for identifying one or more markers of a disease or condition comprising: a) determining a first profile of analytes from a sample comprising a cell-free bodily fluid from a subject having said disease or condition and a population of phagocytic cells, or a population of >2n phagocytic cells, from a subject having said disease or condition; determining a second profile of analytes from a population of =2n phagocytic cells, or a population of non-phagocytic cells, from the subject having said disease or condition; identifying a set of differences between the first and second profiles, wherein the first set of differences is specific to the first profile relative to the second profile; b) determining a third profile of analytes from a sample comprising a cell-free bodily fluid from a subject having said disease or condition and a population of phagocytic cells, or a population of >2n phagocytic cells, from a control subject not having said disease or condition; determining a fourth profile of analytes from a population of =2n phagocytic cells, or a population of non-phagocytic cells, from the control subject not having said disease or condition; identifying a set of differences between the third and

fourth profiles, wherein the second set of differences is specific to the third profile relative to the fourth profile; and c) identifying one or more analytes specific to the set of differences identified in a) relative to the set of differences identified in b), the identified analytes in c) being markers of said disease or condition. In some embodiments, this invention provides a method for identifying one or more markers of a disease or condition comprising: a) determining a first profile of analytes from a sample comprising a cell-free bodily fluid from a subject having said disease or condition and a population of phagocytic cells, or a population of >2n phagocytic cells, from a subject having said disease or condition; b) comparing the first profile to a second profile derived from a repository of analytes from a control subject not having said disease or condition; c) identifying a set of differences between the first and second profiles, wherein the set of differences is specific to the first profile relative to the second profile; and d) identifying one or more analytes specific to the set of differences, the identified analytes being markers of said disease or condition. In further embodiments, the method further comprises: e) obtaining a fifth profile of analytes from cells or tissues affected by said disease or condition in the subject having said disease or condition; obtaining a sixth profile of analytes from cells or tissues not affected by said disease or condition in the subject having said disease or condition; identifying a set of differences between the fifth and sixth profiles, wherein the set of differences is specific to the fifth profile relative to the sixth profile; and f) identifying at least one of the one or more markers of c) present in the set of differences identified in d). In some embodiments, this invention provides a method for identifying one or more markers that may be used in the treatment of a disease or condition. For example, a marker (e.g., protein or gene) identified by a method of the invention may be used as a molecular target for a therapeutic agent. A marker identified by a method of the invention also may be used in any of the other methods of the invention, e.g., for monitoring the progression or regression of a disease or condition. In certain embodiments, the one or more markers identified by the methods of this invention may have therapeutic potential. For example, if a marker is identified as being up-regulated (or down-regulated) in circulating diseased cells from a subject having a disease or condition , a compound or an agent that is capable of down-regulating (up-regulating) said marker may be useful in treating said disease or

condition. Similarly, a gene/protein/lipid/carbohydrate expression profile, a single nucleotide polymorphism profile, a gene mutation profile, a gene copy number profile, a DNA methylation profile, a DNA acetylation profile, a chromosome dosage profile, a gene expression profile, or a combination thereof may be useful in these embodiments.

[0134] An exemplary method for detecting the presence or absence of an analyte (*e.g.*, DNA, RNA, protein, polypeptide, carbohydrate, lipid or the like) corresponding to a marker of the invention in a biological sample involves obtaining a bodily fluid sample (*e.g.*, blood) from a test subject and contacting the bodily fluid sample with a compound or an agent capable of detecting one or more markers. Detection methods described herein can be used to detect one or more markers in a biological sample *in vitro* as well as *in vivo*. For example, *in vitro* techniques for detection of mRNA include Northern hybridizations and in situ hybridizations. *In vitro* techniques for detection of a polypeptide corresponding to a marker of the invention include enzyme linked immunosorbent assays (ELISAs), Western blots, immunoprecipitations and immunofluorescence. *In vitro* techniques for detection of genomic DNA include Southern hybridizations. Furthermore, *in vivo* techniques for detection of a polypeptide corresponding to a marker of the invention include introducing into a subject a labeled antibody directed against the polypeptide. For example, the antibody can be labeled with a radioactive marker whose presence and location in a subject can be detected by standard imaging techniques. Because each marker is also an analyte, any method described herein to detect the presence or absence of a marker can also be used to detect the presence or absence of an analyte.

[0135] The marker that is useful in the methods of the invention can include any mutation in any one of the above-identified markers. Mutation sites and sequences can be identified, for example, by databases or repositories of such information, *e.g.*, The Human Gene Mutation Database (www.hgmd.cf.ac.uk), the Single Nucleotide Polymorphism Database (dbSNP, www.ncbi.nlm.nih.gov/projects/SNP), and the Online Mendelian Inheritance in Man (OMIM) website (www.ncbi.nlm.nih.gov/omim).

[0136] A marker that is useful in the methods of the invention can include any marker that is known to be associated with a disease or condition. Markers that can be used in this invention can be any marker that has been well-characterized as associated with a specific disease or condition, or any markers that have bee identified by the methods of this invention.

[0137] In some embodiments, the markers comprise at least one gene selected from the group consisting of AKT2, BAK1, EGFR, ERBB2, ETS2, FOS, JUN, MAP2K1, MMP2, PDGFB, RB1, SERPINB2, SNCG, and SPP1. In some embodiments, the one or more markers comprise at least one gene selected from the group consisting of AKT1, AKT2, BAK2, CDC25A, E2F1, EGFR, ERBB2, FOS, JUN, MAP2K1, MMP2, NFKB1, PDGFB, PIK3R1, PNN, RB1, SERPINB2, SERPINB5, SNCG, SPP1, TERT, TIMP3, and TP53. In some embodiments, the one or more markers comprise at least one gene selected from the group consisting of CASP8, CASP9, COL18A1, ETS2, HTATIP2, MMP9, SRC, and TWIST1. In some embodiments, the one or more markers comprise at least one gene selected from the group consisting of AKT1, APAF1, ATM, CDC25A, CDKN1A, ETS2, FOS, IL8, ITGA4, ITGA6, ITGAV, JUN, MAP2K1, NFKBIA, PLAU, PLAUR, RAF1, SERPINB2, SYK, TIMP1, TNF, TNFRSF10B, and TNFRSF1A. In some embodiments, the markers comprise at least one gene selected from the group consisting of ACP2, AK2, AKT3, ARL5B, ATP2B3, BGN, BRAF, BTG2, CAMKK2, CAPG, CAPN12, CPLX2, DENND5A, DNA2, FAM104A, FNIP1, GFRA4, GLUD1, GNAQ, GP1BB, HNRPLL, HOXA2, HPS3, INPP4A, ITGAV, KLHL23, LANCL2, LYPD6, MAPKAPK3, MEF2A (includes, EG:4205), MEF2C, NVL, PCYT1A, PGLYRP4, PLOD1, PPP1CB, PRKAB2, PROS1, PTPRE, RASA4 (includes,EG:10156), RBMS2, RBPJ, STAT5B, THBS1, TRIB1, TRIM2, TSPAN6, and ZDHHC21. In some embodiments, the markers comprise at least one gene selected from the group consisting of B4GALT5, BOP1, CCL2, CCL3, CCL3L1, CCRL2, CD83, CLEC4G, CLIC4, CTSC, CTSO, CXCL10, FCGR3A, FPR3, HBA1, HBB, LRMP, MAP1LC3B2, MS4A4A, MSR1, MYADML, NID1, PF4, PION, RNF217, SAMD9L, SERPING1, and SPARC. In some embodiments, the markers comprise at least one gene selected from the group consisting of ACOT9, AMPD2, ARHGAP15, BATF2, C3AR1, C5orf41, CCL3, CCL3L1, CD63, CHST11, CHSY1, CLEC4G, CTSZ,

CXorf21, CYTH4, CYTIP, DLEU2, DNAJA1, DOCK8, DTX3L, DUSP6, EPSTI1, ERF, F2RL1, FYB, GABRB2, GBP5, GLRX, GNB4, ICAM1, IFI35, IFIH1, IFNAR2, IL1R1, IRF1, ITGA5, LAP3, LAPTM5, LCP2, MAP1LC3B, MAP1LC3B2, MICAL2, MT1DP, MT1JP, MT1M, MT2A, MYADML, NEK6, NINJ2, NNMT, NT5C3L, NUB1, PDE4B, PLOD1, PML, PRKCB, PSMB9, RCN3, RGS4, RNASE6, RTP4, SAMD9L, SEL1L, SERPING1, SETX, SIGLEC10, SKIL, SLC7A7, SNORA21, SP100, SP110, SP140, SSFA2, STAT2, STK17B, STK3, TDRD7, TMCC1, TMPRSS11E2, TNFRSF1B, TPM1, TRIM21,TXNDC4, UBE2L6, UBE2W, USP18, VAV1, WARS, WIPF1, and WIPI1. In some embodiments, the markers comprise at least one gene selected from the group consisting of ADAR, ADM, ALAS1, ANKRD22, ARHGAP27, B3GNT5, BCL10, C12orf35, C15orf29, C2orf59, CD177, CEACAM1, CPEB2, DDX58, F2RL1, GDPD3, GNAI3, HIST2H3A, HIST2H3D, HIST2H4A, HMGCR, HSPA6, HSPC159, IL4R, IMPA2, KPNB1, KREMEN1, KRT23, LDLR, LOC100130904, LTB4R, MAEA, MARK2, MBOAT2, MPZL3, N4BP1, NBEAL2, NMI, NPEPPS, PARP14, PGM2, PPIF, PXN, RALBP1, ROD1, RPS6KA1, S100P, SERTAD2, SLC9A1, SLPI, SP110, SPINT1, ST14, TBC1D3, TNFRSF9, TRIM21, UPP1, VPS24, ZBTB34, and ZNF256.

[0138] In some embodiments, the markers comprise at least one biomarker selected from the group consisting of ACTN4, BC020163, CMIP, CNN2, EDNRB, GPM6B, KIT, MGC40222, NAMPT, PRAME, RPL18, RPL21, RPS15, TMEM80, TRIB2, TTC3, and VDAC1. In one embodiment, the markers are ACTN4, BC020163, CMIP, CNN2, EDNRB, GPM6B, KIT, MGC40222, NAMPT, PRAME, RPL18, RPL21, RPS15, TMEM80, TRIB2, TTC3, and VDAC1. These markers are useful in the diagnosis, prognosis, or monitoring of melanoma, or discriminating between different types of skin lesions, for example, melanoma and naevi (See, e.g., Wachsman et al., "Noninvasive genomic detection of melanoma," *Br J Dermatol.* 2011 Apr;164(4):797-806).

[0139] In some embodiments, the marker that is useful in the methods of the invention for prenatal or pregnancy-related diseases or conditions include those disclosed in, for example, United States Patents 7,655,399, 7,651,838, 6,660,477, 6,172,198, 5,594,637, 5,514,598, 6,258,540, 6,664,056, 7,235,359, and 7,645,576,

United States Patent Application Publications 20090162842, 20090155776, 20070207466, 20060019278, 20040086864, 20020045176, 20010051341, 20020192642, 20040009518, 20040203037, 20050282185, 20060252071, 20070275402, 20080153090, 20090170102, 20090061425, 20020045176, 20040137452, 20050164241, 20060019278, 20060252068, 20060252071, 20060257901, 20070141625, 20070218469, 20070275402, 20090155776, 20090162842, 20090170102, 20090317797, 20100120056, 20100120076, and 20100137263 and International Patent Application Publications WO/2006/026020, WO/2002/068685, WO/2005/111626, WO/2009/055487, WO/2009/001392, and WO/2008/014516.

[0140] In some embodiments, the marker that is useful in the methods of the invention for neurological or neuropsychiatric diseases or conditions include those disclosed in, for example in United States Patents 7,723,117, 6,867,236, United States Patent Application Publications 20060115854, 20060115855, 20060166283, 20060234301, 20060259990, 20060259991, 20070162983, 20070264197, 20080026405, 20080038730, 20080051334, 20080152589, 20080220013, 20080261226, 20080269103, 20080286263, 20090041862, 20090239241, 20090275046, 20090318354, 20090324611, 20100009352, 20100021929, 20100028356, 20100055722, 20100062463, 20100075891, 20100105623, 20100124756, 20100159486, 20100167937, 20100169988, 20100167320, 20100112587, 20100098705, 20100068705, 20100009356, 20090305265, 20100124746, 20100092983, 20070148661, 20070141625, 20100120050, 20090155230, 20090274709, International Patent Application Publications WO/2004/040016, WO/2004/071269, WO/2005/033341, WO/2005/052592, WO/2005/103712, WO/2005/114222, WO/2006/020269, WO/2006/048778, WO/2006/050475, WO/2006/061609, WO/2006/105907, WO/2006/133423, WO/2006/134390, WO/2007/098585, WO/2007/119179, WO/2008/010660, WO/2008/014314, WO/2008/028257, WO/2008/046509, WO/2008/046510, WO/2008/046511, WO/2008/046512, WO/2008/063369, WO/2008/085035, WO/2008/095261, WO/2008/100596, WO/2008/120684, WO/2008/125651, WO/2008/127317, WO/2008/132464, WO/2009/000520, WO/2009/001392, WO/2009/068591, WO/2009/074331, WO/2009/100131, WO/2010/005750,

WO/2010/011506, WO/2010/019553, WO/2010/059242, WO/2010/061283, WO/2010/063009, WO/2010/066000, WO/2009/121152, WO/2009/121951, WO/2009/097450, WO/2009/092382, WO/2009/075579, WO/2009/058168, WO/2009/053523, WO/2009/034470, WO/2009/032722, WO/2009/014639, WO/2009/003142, WO/2010/041046, WO/2007/131345, WO/2008/003826, and WO/2009/07556.

[0141] In some embodiments, the marker that is useful in the methods of the invention for cardiovascular diseases or conditions include those disclosed in, for example in United States Patents 7,670,769, 7,445,886, 7,432,107, 7,157,235, and 7,009,038, United States Patent Application Publications 20100167320, 20100112587, 20100098705, 20100068705, 20100009356, 20090305265, 20100124746, 20100092983, 20070148661, 20070141625, 20100120050, 20090155230, and 20090274709, and International Patent Application Publications WO/2009/121152, WO/2009/121951, WO/2009/097450, WO/2009/092382, WO/2009/075579, WO/2009/058168, WO/2009/053523, WO/2009/034470, WO/2009/032722, WO/2009/014639, WO/2009/003142, WO/2010/041046, WO/2007/131345, WO/2008/003826, and WO/2009/075566.

[0142] In some embodiments, the marker that is useful in the methods of the invention for kidney-associated diseases or conditions include those disclosed in, for example in United States Patents 7,488,584, 7,459,280, 7,294,465, and 7,662,578, United States Patent Application Publications 20100143951, 20100124746, 20100120056, 20100120041, 20100081142, 20090155230, and 20090239242, International Patent Application Publications WO/2010/059996, WO/2010/054389, WO/2010/048347, WO/2010/048497, WO/2010/054167, WO/2010/048346, WO/2010/046137, WO/2010/025434, WO/2010/018185, WO/2010/012306, WO/2009/122387, WO/2009/083950, WO/2009/080780, WO/2009/060035, WO/2009/059259, WO/2008/154238, WO/2008/089936, WO/2008/084331, WO/2008/042012, WO/2007/131345, WO/2005/012907, WO/2004/024098, WO/2003/019193, WO/2007/112999, WO/2007/082733, WO/2006/073941, WO/2010/068686, WO/2010/022210, and WO/2009/127644.

[0143] In some embodiments, the marker that is useful in the methods of the invention for autoimmune or immune-related diseases or conditions include those disclosed in, for example 7,604,948, 7,670,764, 6,986,995, and 6,631,330, United States Patent Application Publication 20070141625, 20090263474, 20100075891, 20100104579, 20100105086, 20100131286, 20090176217, 20090202469, 20020119118, 20090258025, 20100137393, 20100120629, 20090318392, 20090196927, 20090023166, 20080227709, 20080039402, 20080026378, 20070224638, 20070218519, 20060210562, 20050266432, 20050164233, 20050130245, 20090130683, 20090110667, 20090054321, 20090023166, and 20080274118, and International Patent Application Publication WO/2009/043848, WO/2010/053587, WO/2010/046503, WO/2010/039714, WO/2009/100342, WO/2009/053537, WO/2009/017444, WO/2008/156867, WO/2008/147938, WO/2008/129296, WO/2008/137835, WO/2008/082519, WO/2008/064336, WO/2008/043782, WO/2008/043725, WO/2007/047907, WO/2006/125117, WO/2006/114661, WO/2006/020899, WO/2005/114222, WO/2005/007836, WO/2004/076639, WO/2004/050704, and WO/2001/014881.

[0144] The present invention also provides kits that comprise marker detection agents that detect at least one or more of the markers identified by the methods of this invention. This present invention also provides methods of treating or preventing a disease or condition in a subject comprising administering to said subject an agent that modulates the activity or expression or disrupts the function of at least one or more of the markers identified by the methods of this invention.

[0145] It is to be understood that the embodiments of the present invention which have been described are merely illustrative of some of the applications of the principles of the present invention. Numerous modifications may be made by those skilled in the art based upon the teachings presented herein without departing from the true spirit and scope of the invention.

[0146] The following examples are set forth as being representative of the present invention. These examples are not to be construed as limiting the scope of the

invention as these and other equivalent embodiments will be apparent in view of the present disclosure and accompanying claims.

Examples

Representative Method I for the Separation of Phagocytic Cells with DNA Content of 2n from Non-Phagocytic Cells and the Analysis of Expression Profiles

[0147] 1. Separate blood sample into plasma and buffy coat including WBC sample. Coat plates to receive WBC sample with avidin.

[0148] 2. Add biotinylated antibody to non-phagocytic blood cell (e.g., T cells) to the wells, incubate for 30 min at RT, wash wells.

[0149] 3. Add magnetic beads.

[0150] 4. Add WBC blood sample.

[0151] 5. Incubate at 37 °C (30 minutes – 1 hour).

[0152] 6. Following phagocytosis of beads by phagocytic cells and binding of avidin-biotin-antibody to non-phagocytic cells, place plate on top of magnet and wash (the phagocytic cells that internalized the magnetic beads and the non-phagocytic cells bound to the antibody will stay; all other cells will be washed away).

[0153] 7. Remove magnet and collect phagocytic cells and separate into phagocytic cells with DNA equal to 2n and DNA greater than 2n. Non-phagocytes and phagocytes having DNA equal to 2n are referred to as cells having DNA equal to 2n.

EXAMPLE 2

Representative Method II for the Separation of Phagocytic Cells from Non-Phagocytic Cells

[0154] 1. Separate blood sample into plasma and buffy coat including WBC sample.

[0155] 2. Cytospin WBC on glass slides.

[0156] 3. Fix cells in acetone/methanol (-20 °C for 5 minutes).

[0157] 4. Stain with hematoxylin and eosin stain and anti-T cell antibody.

[0158] 5. Isolate T cells (non-phagocytic) and macrophages (phagocytic) using laser capture microscopy (LCM). Separate into phagocytic cells with DNA equal to 2n and DNA greater than 2n. Non-phagocytes and phagocytes having DNA equal to 2n are referred to as cells having DNA equal to 2n.

EXAMPLE 3

Representative Method III for the Separation of Phagocytic Cells from Non-Phagocytic Cells

[0159] 1. Separate plasma from whole blood.

[0160] 2. Use magnetic antibody-conjugated beads to isolate non-phagocytic (e.g., T cells) and phagocytic cells (e.g., neutrophils and/or macrophages and/or monocytes) from whole blood. Separate into phagocytic cells with DNA equal to 2n and DNA greater than 2n. Non-phagocytes and phagocytes having DNA equal to 2n are referred to as cells having DNA equal to 2n.

EXAMPLE 4

Representative Method IV for the Separation of Phagocytic Cells from Non-Phagocytic Cells and the Analysis of Expression Profiles

[0161] 1. Separate blood sample into plasma and buffy coat including WBC sample. Stain WBC with fluorescent antibodies specific against a particular cell subpopulation (e.g., neutrophils, macrophages, monocytes, T cells and the like) and a DNA stain, (e.g., Hoechst 33342, Propidium iodide).

[0162] 2. Sort the cells (e.g., by FACS).

EXAMPLE 5

Representative Method for the Analysis of Expression Profiles

[0163] 1. Create a combination sample by combining two or more different components selected from: a cell-free bodily fluid isolated from a subject, phagocytic cells isolated from the subject, >2n phagocytic cells isolated from the subject, circulating vesicles isolated from the subject, and circulating diseased cells isolated from the subject.

[0164] 2. Isolate RNA from the combination sample and from a control sample with a component selected from: =2n phagocytic cells isolated from the subject, non-phagocytic cells isolated from the subject, and control cells isolated from the subject. Prepare cDNA or cRNA and use to differentiate genetic profiles (e.g., a cancer gene array) between the combination sample and the control sample.

[0165] 3. Isolate DNA from the combination sample and from the control sample. Run DNA arrays and compare the profiles obtained from the combination sample and the control sample.

[0166] 4. Isolate protein from the combination sample and from the control sample. Run Western blots using antibodies to known proteins overexpressed by human tumors (e.g., PSA and PSMA in prostate cancer; CEA in colon cancer; and CA125 in ovarian cancer), and compare the profiles obtained from the combination sample and the control sample.

[0167] 5. Isolate lipids from the combination sample and from the control sample. Compare quantity and quality of lipids, for example using HPLC, between the combination sample and the control sample.

EXAMPLE 6

[0168] A fraction of pregnancies are complicated by the presence of genetic abnormalities in the fetus, and considerable clinical interest has focused on development of assays to be able to detect and characterize these abnormalities. Among the most common problems is the presence of chromosomal aneuploidies in which one or more of the chromosomes is present at other than the normal

abundance. The incidence of aneuploidies increase in frequency with maternal age and encompass potentially viable pregnancies with deviations from diploid number of chromosomes 13, 18 and 21, or abnormal numbers of X chromosomes in either male or female fetuses. Detection of these aneuploid conditions can occur by direct analysis of fetal cells collected by amniocentesis or chorionic villus sampling (CVS) and using traditional cytogenetic karyotyping or array comparative genome hybridization (CGH), quantitative PCR (qPCR) or sequencing of fetal genomic DNA, but these approaches carry a risk of fetal or maternal morbidity and miscarriage as a consequence of the trauma of cell collection. Additionally, these invasive cell collection procedures are only effective when applied late in the first trimester of pregnancies, and earlier genotyping is preferred to enable earlier and better decisions on at risk pregnancies.

[0169] In order to avoid the morbidity risks of invasive testing, non-invasive procedures have been developed to detect fetal aneuploidies by analyzing the sequence composition of degraded genomic fetal DNA present at low levels in maternal blood, also known as circulating cell-free fetal DNA (cfff DNA). However, these procedures rely on the exclusion of intact maternal genomic DNA from cellular components of blood to increase the observable signals that the procedures rely upon. By contrast, other methods use purely cellular fractions, but these methods may have high levels of maternal cell contamination (Bianchi et al., Proc Natl Acad Sci USA. 87(9):3279–3283 (1990); Bianchi et al., Am J Hum Genet. 61(4):822–829 (1997)).

[0170] In contrast to the methods mentioned above, and as proof-of-principle of the methods of the invention, an exemplary method was developed that harnesses the power of procedures reliant on cfff DNA and procedures reliant on cellular fractions alone. Here, the approach explicitly includes a cell-free bodily fluid and phagocytic blood cells expected to contain fetal genomic markers scavenged by phagocytosis of fetal cells in the placenta and other compartments. Evidence of fetal DNA existing in a maternal sample is provided by the presence of paternal markers, which are DNA sequences that do not exist in the mother and are novel when comparing a strictly maternal DNA sample to a combination sample of maternal and fetal DNA. In

practice, paternal markers may be seen by DNA sequencing analyses in selected segments of the human genome which contain polymorphic sites. Polymorphic sites are often single nucleotide polymorphisms (SNPs), and distinguish alleles. Markers of this type are informative when the mother is homozygous for a SNP sequence and the father has a variant SNP passed to the fetus. For example, if the mother carries A/A nucleotides at a particular SNP location, and the father carries G/G nucleotides at that same SNP location, the G allele of the SNP will be found at a proportion indicating the relative amount of fetal DNA in a sample. Using sequencing techniques, the proportion of fetal material may be calculated by comparing the number of "A" bearing molecules (the maternal-derived material) to the number of "G" molecules (the fetal-derived material). Exemplary assays for calculating fetal DNA content in a sample using this rationale are described in, for example, Chiu et al., Proc Natl Acad Sci USA 105:20458–63 (2008); Ehrich et al., Am J Obstet Gynecol 204:205 e201–211 (2011); and Sehnert et al., Clin Chem 57:1042–9 (2011), each of which is herein incorporated by reference in its entirety. Calculation of fetal DNA content is accurate, as exemplified by Tynan et al. (J Mol Diagn 13(4):382-9 (2011)) in a study indicating that the confidence level of correctly detecting the presence of fetal DNA in a combination sample with 6 or more informative SNPs was greater than 99.9%. Where the mother is homozygous, the fetal fraction is seen directly as the variant fraction, while if the mother is heterozygous the fetal allelic fraction is added to the maternal, and may distort the observed ratio if the fetus is homozygous. These cases undoubtedly exist in the data shown, but have not been analyzed to simplify the analysis.

[0171] To produce an unbiased panel of potentially informative SNPs, criteria were selected such that all populations should carry the reference and alternative SNPs in similar proportions and the minor allele frequency (MAF) should be greater than 40%. Catalogs of SNPs having these qualities are commercially available (Durbin et al., Nature 467(7319):1061–1073 (2010) and McVean et al., Nature 491:56–65 (2012)). MAFs are the frequencies at which the least common allele for a given gene occurs in a population. MAFs of greater than 40% were targeted to increase the likelihood of finding heterozygotes in the fetal fraction, while at the same time having abundant homozygotes in the maternal fraction, thus increasing the likelihood of

finding informative SNPs. An initial set of 167 markers was chosen for analysis. Based on previous analyses, for any mother-fetus combination sample (e.g., a cell-free bodily fluid in combination with phagocytes), 10-20% of the panel SNPs would be expected to be informative. PCR primers were designed that contained the 167 SNP markers and were of a small enough size (80-120 bp) so that the probability of amplification from partially degraded DNA would be greatest. The list of SNPs and their genomic references and the primer sequences used to amplify them are provided in Table 1.

Table 1: SNPs and genomic references

GenBank Target ID	Target Interval	Left Primer With Tag	Right Primer With Tag
rs2937436.G.A	chr5:31325718-31325728	ACACTGACGACATGGTTCTACAAAAGAAAGGTATCATCTGAAGTT	TACGGTAGCAGAGACTTGGTCTTAGAACTACTTACTTTTACTTATTAGGA
rs6946634.A.G	chr7:79093092-79093102	ACACTGACGACATGGTTCTACAGGATTTTAAAATGAAAATTGAAGAAGTA	TACGGTAGCAGAGACTTGGTCTTGTTGTTCTTTGGTCTGTAAAAT
rs1060922.T.C	chr18:722398-722408	ACACTGACGACATGGTTCTACATTAGTACCAGAACCACTGC	TACGGTAGCAGAGACTTGGTCTATTTACCTGATTTTAATCATCAGAT
rs11734833.A.G	chr4:101108065-101108075	ACACTGACGACATGGTTCTACAACAGGTATTCATCATTCACTC	TACGGTAGCAGAGACTTGGTCTGTTTCCCCTGTTCTTAAGTG
rs10490392.G.T	chr2:186658561-186658571	ACACTGACGACATGGTTCTACAATCAGCTTATGCTGATGATAATC	TACGGTAGCAGAGACTTGGTCTAAAATAGTTTTAGTAATGAAGTTAGCA

GenBank Target ID	Target Interval	Left Primer With Tag	Right Primer With Tag
rs1872575.G.A	chr3:113804975-113804985	ACACTGACGACATGGTTCTACATTAAATGACCTGGATTGATCAG	TACGGTAGCAGAGACTTGGTCTCTTTTGTCTTTGACTAAATGAATCT
rs4241779.A.G	chr4:184600597-184600607	ACACTGACGACATGGTTCTACAATTGGATGCAATTGCTCAG	TACGGTAGCAGAGACTTGGTCTACTGTTTAGACATCCATGCA
rs1958716.A.G	chr14:20528317-20528327	ACACTGACGACATGGTTCTACAGGAACTTCAAATTTTCTTCTTTG	TACGGTAGCAGAGACTTGGTCTAGGGTTGACCTACATGTC
rs3085724.TCA.T	chr3:156255511-156255523	ACACTGACGACATGGTTCTACATTTGAGCTTTACAAATAAACATACA	TACGGTAGCAGAGACTTGGTCTATCCCAGTGCTAATTAAACAA
rs5880678.TC.T	chr6:146756123-146756134	ACACTGACGACATGGTTCTACAGCTGCCTTAAGTAGGAAGA	TACGGTAGCAGAGACTTGGTCTGTGCTTTGTCAGAATTCCC
rs10741130.T.C	chr10:27475440-27475450	ACACTGACGACATGGTTCTACAGAAACAGATACCTCCTATTTTGA	TACGGTAGCAGAGACTTGGTCTCAAGGCTAGACTAAAAGGAAAA
rs2305641.G.A	chr12:69646910-69646920	ACACTGACGACATGGTTCTACATTCATTCTTTGGGAGTAAATGA	TACGGTAGCAGAGACTTGGTCTCTGTATCACTAGCTACTAAATCAAAA
rs104220	chr3:151546037-	ACACTGACGACATGGTTCTACAA	TACGGTAGCAGAGACTTGGTCTA

GenBank Target ID	Target Interval	Left Primer With Tag	Right Primer With Tag
1.G.A	151546047	AAACCTCAGAAAATTTGCATT	AAACAGAAAAGTGCCTATTTAAG
rs3806.G.A	chr13:103330714-103330724	ACACTGACGACATGGTTCTACAGCGTATTCTAAAATAGGAAACAAG	TACGGTAGCAGAGACTTGGTCTATAGTGGATGAAAACATGCATT
rs9307834.A.C	chr4:76453846-76453856	ACACTGACGACATGGTTCTACAGTATTTAAGAATTTTCTGTTTTAATGC	TACGGTAGCAGAGACTTGGTCTGTGGGTTAAATCCCATTTAGA
rs6919254.T.C	chr6:136599400-136599410	ACACTGACGACATGGTTCTACACATAAGCTGAAAGGCCAG	TACGGTAGCAGAGACTTGGTCTATGACCCATCTGAGTCTATC
rs1338.T.C	chr10:27475689-27475699	ACACTGACGACATGGTTCTACAAGTTAAAAGGCTGAAAGGAATA	TACGGTAGCAGAGACTTGGTCTAGTTTTCATAAAGTGCAATAAACA
rs2999278.C.T	chr10:44053581-44053591	ACACTGACGACATGGTTCTACAAGGAGGAAAGCTCATGTAA	TACGGTAGCAGAGACTTGGTCTAAGTCTACCCTTTGTTCTTAC
rs60029104.G.A	chr2:186658052-186658062	ACACTGACGACATGGTTCTACATATCACGTAAAGGCAAATGT	TACGGTAGCAGAGACTTGGTCTTGAAGTTGAAGTACTTTTCGATA
rs1862066.G.A	chr2:186671908-186671919	ACACTGACGACATGGTTCTACAAGAAAATTTAGCAAGAAGACTAAC	TACGGTAGCAGAGACTTGGTCTTGAAACTGAAACCTCATCAC

GenBank Target ID	Target Interval	Left Primer With Tag	Right Primer With Tag
	18		
rs4518168.G.A	chr3:97806612-97806622	ACACTGACGACATGGTTCTACAGTTCAAAATTTCATGCAATGGT	TACGGTAGCAGAGACTTGGTCTAGTCATTAAAGTGGGTATTTGTA
rs10816767.G.A	chr9:111782666-111782676	ACACTGACGACATGGTTCTACAGCATTCTGTATGTAAGGCC	TACGGTAGCAGAGACTTGGTCTGGAAGTTCAAATTTGACAGC
rs7138345.G.A	chr12:104290853-104290863	ACACTGACGACATGGTTCTACATAAACTGATGGAAGCAATGAAT	TACGGTAGCAGAGACTTGGTCTGTAGGTACTCAGCTTGAAAC
rs6917033.G.A	chr6:42627430-42627440	ACACTGACGACATGGTTCTACACATAACCTTTGTATTGCAGCC	TACGGTAGCAGAGACTTGGTCTATACATGTAACGAATTGTCTTTG
rs115032519.T.C	chr6:30002877-30002887	ACACTGACGACATGGTTCTACACTATTGCCTGGATCCCAT	TACGGTAGCAGAGACTTGGTCTAAGCTTAGGCTTTCAGATTTTAC
rs3747562.A.G	chr12:42854201-42854211	ACACTGACGACATGGTTCTACACCGGATTTCAATGTCATAGTTC	TACGGTAGCAGAGACTTGGTCTGTACATCTGCCAGTGCTC
rs2240859.A.G	chr3:42787465-42787475	ACACTGACGACATGGTTCTACAGAATTTGTTGAGCCCGAC	TACGGTAGCAGAGACTTGGTCTTATTTCAGGTTTTGGCCAGA

GenBank Target ID	Target Interval	Left Primer With Tag	Right Primer With Tag
	5		
rs1050287.T.C	chr12:27954503-27954513	ACACTGACGACATGGTTCTACAAGCATAGCAATGTAAAAGGAAT	TACGGTAGCAGAGACTTGGTCTACTCATTCATAATGCCTCCTTA
rs1573612.T.C	chr12:12047425-12047435	ACACTGACGACATGGTTCTACAGCATCACCAATGTTCCAA	TACGGTAGCAGAGACTTGGTCTGAGGGTTCCAAACATACTG
rs1230381.A.G	chr5:96254205-96254215	ACACTGACGACATGGTTCTACATTATCAAGAAGGCCAGGAA	TACGGTAGCAGAGACTTGGTCTGACCATTCATTTGGAATGCT
rs9880307.C.A	chr3:16974606-16974616	ACACTGACGACATGGTTCTACATCTAAGAAAATGAAACGTACCT	TACGGTAGCAGAGACTTGGTCTGTCCTTCACTCAAATTTCCT
rs3800989.A.G	chr7:141360581-141360591	ACACTGACGACATGGTTCTACAAACCATTACTTGACAGAAAGG	TACGGTAGCAGAGACTTGGTCTGTTGAAGCCTCACCTTTTAT
rs28394363.G.A	chr3:33428295-33428305	ACACTGACGACATGGTTCTACAATCTGATAAAATGACCCTACTC	TACGGTAGCAGAGACTTGGTCTCACCTAAGTTTACAATGAAGG
rs7337667.G.A	chr13:97489453-97489463	ACACTGACGACATGGTTCTACAGGAAATTAATACTGATAAAGACTATCAG	TACGGTAGCAGAGACTTGGTCTTTTAGTCTTATCTTTCGCAAAACATA

GenBank Target ID	Target Interval	Left Primer With Tag	Right Primer With Tag
rs58050565.ATTT.A	chr9:88937847-88937860	ACACTGACGACATGGTTCTACATGAAGATAAGTTTACTGAATTCATAAA	TACGGTAGCAGAGACTTGGTCTCTCCTAGTGGAACTTAATAAAATAAGT
rs3734110.T.C	chr5:13701532-13701542	ACACTGACGACATGGTTCTACAAGTTCAGTAAACCAGAAACC	TACGGTAGCAGAGACTTGGTCTGTGAACATAGCTGGTTTACA
rs5930933.C.T	chrX:135431354-135431364	ACACTGACGACATGGTTCTACATACCCCTGTTTCATACCC	TACGGTAGCAGAGACTTGGTCTCTCAGCTTCAGTACTATGAGG
rs831618.T.G	chr11:33728610-33728620	ACACTGACGACATGGTTCTACAGATCCAGCTAACTTTGCATA	TACGGTAGCAGAGACTTGGTCTTGGTACAGGATTCTAAGCT
rs2104772.T.A	chr9:117808781-117808791	ACACTGACGACATGGTTCTACACAATTTCGTATTCAGTAGCCT	TACGGTAGCAGAGACTTGGTCTGTCTGAGCCACTGGAAATA
rs2269991.A.G	chr7:79090132-79090142	ACACTGACGACATGGTTCTACACCTTTTGTAATTGTAGGCATTC	TACGGTAGCAGAGACTTGGTCTGGGGATAGATTCAGCTACT
rs3766601.C.T	chr1:100388231-100388241	ACACTGACGACATGGTTCTACATGATAACTTATTTTAAGATGCCTTAAAA	TACGGTAGCAGAGACTTGGTCTAAATTCTATCTAGGAAATAAGTTCTATGA
rs428463	chr16:70	ACACTGACGACATGGTTCTACAT	TACGGTAGCAGAGACTTGGTCTT

GenBank Target ID	Target Interval	Left Primer With Tag	Right Primer With Tag
2.C.G	606551-70606561	AGTATGGGGATTTTAGCATG	CTTCTGTACTGACCCTTC
rs9389248.T.C	chr6:135282652-135282662	ACACTGACGACATGGTTCTACATAACATCAGATGAGGTCACT	TACGGTAGCAGAGACTTGGTCTGGTTGGCTGTTTATGGTTT
rs1054975.T.C	chr3:9406832-9406842	ACACTGACGACATGGTTCTACACCACTAACCAACTCCTAGAT	TACGGTAGCAGAGACTTGGTCTACAGTGGCTCCAATAGTG
rs6438874.G.A	chr3:124739888-124739898	ACACTGACGACATGGTTCTACACAGACTGACCGTCAAAGA	TACGGTAGCAGAGACTTGGTCTAAGCTTAACAACTCCACTG
rs832068.G.A	chr3:97754143-97754153	ACACTGACGACATGGTTCTACATTTTCCGGGGAAGGTTTT	TACGGTAGCAGAGACTTGGTCTACCTCACTTCTTTGTTTGC
rs840388.T.G	chr5:34018886-34018896	ACACTGACGACATGGTTCTACAACTGGAAAATTACCAAGGC	TACGGTAGCAGAGACTTGGTCTCCGCAAATTCTAAATCTTAAAACA
rs7023954.G.A	chr9:21816754-21816764	ACACTGACGACATGGTTCTACATGATGCCTTAATTTTGGGG	TACGGTAGCAGAGACTTGGTCTCCAGAATTTAAGTTAAATTTAAATTATCCT
rs7808823.T.C	chr7:129989874-129989898	ACACTGACGACATGGTTCTACAGAACAAGGTGGACTTCTG	TACGGTAGCAGAGACTTGGTCTGCTGTAAGCAAGTCCATGA

GenBank Target ID	Target Interval	Left Primer With Tag	Right Primer With Tag
	84		
rs35350755.T.G	chr10:123549687-123549697	ACACTGACGACATGGTTCTACAAAACAAAGCCAATGATTTCC	TACGGTAGCAGAGACTTGGTCTGAAAACTCCAAGTACTGCC
rs3802980.A.G	chr11:5706307-5706317	ACACTGACGACATGGTTCTACAGGTTAGGATTCACTCACCA	TACGGTAGCAGAGACTTGGTCTATTGGCCCATAGTTTATCTTTC
rs2942.G.A	chr6:146755136-146755146	ACACTGACGACATGGTTCTACACACGTGAAGACCAATGAG	TACGGTAGCAGAGACTTGGTCTGTCTTGGTGCTGGTATCT
rs6921.T.A	chr8:103838872-103838882	ACACTGACGACATGGTTCTACACAGACATTTCTTAAAAGTTACCGT	TACGGTAGCAGAGACTTGGTCTCCTTGAACTTCATGGCATA
rs1133181.C.A	chr2:55402268-55402278	ACACTGACGACATGGTTCTACAGAGTTTTCAATCAAATACTTATTCAG	TACGGTAGCAGAGACTTGGTCTGTCGAATTTTGCATAGAAGTAGA
rs3837134.CT.C	chr7:27221061-27221072	ACACTGACGACATGGTTCTACATCTTTCCAAATGTTGCAAGA	TACGGTAGCAGAGACTTGGTCTTTTTCTTTTGCCATTAGTTGAT
rs2414105.A.G	chr15:51868369-51868379	ACACTGACGACATGGTTCTACAAGTCATTTGCCAAAATAACAATA	TACGGTAGCAGAGACTTGGTCTAAATAGTATGTAATAAGATTTCAAAAGT

GenBank Target ID	Target Interval	Left Primer With Tag	Right Primer With Tag
rs7305115.A.G	chr12:72372858-72372868	ACACTGACGACATGGTTCTACACCACTGTACCCAGTACATC	TACGGTAGCAGAGACTTGGTCTGATCATTGACCACTTTATGGT
rs2177142.G.A	chr7:22459320-22459330	ACACTGACGACATGGTTCTACACGGATCCATGTTGACAGA	TACGGTAGCAGAGACTTGGTCTACAGCATTGGATTGCTCA
rs2400478.T.C	chr5:147510862-147510872	ACACTGACGACATGGTTCTACATCTTCTTGAATGCCATAAAGTA	TACGGTAGCAGAGACTTGGTCTGGCTTTAAATCCTTTGGACA
rs6895049.T.G	chr5:52248869-52248879	ACACTGACGACATGGTTCTACAACCAAGAATACATTTAGCATCT	TACGGTAGCAGAGACTTGGTCTCCTGATTTTCCATCCTAGATAA
rs8128.A.C	chr1:115110679-115110689	ACACTGACGACATGGTTCTACAACTTAAACATCAATGGTTTTGG	TACGGTAGCAGAGACTTGGTCTACAATTTAGCAGGTAGAAGAAAA
rs3828977.G.A	chr7:29519494-29519504	ACACTGACGACATGGTTCTACATGACTGGATACCGCTTCT	TACGGTAGCAGAGACTTGGTCTCGGCTTCTGAATAAGGATG
rs12102203.A.G	chr15:51791555-51791565	ACACTGACGACATGGTTCTACATAGATTTAAAGGTCGAATGATCT	TACGGTAGCAGAGACTTGGTCTGAAGCATGCTGTCAAATTT
rs797455	chr12:5023955-	ACACTGACGACATGGTTCTACAC	TACGGTAGCAGAGACTTGGTCTA

GenBank Target ID	Target Interval	Left Primer With Tag	Right Primer With Tag
9.A.T	5023965	ATCTACTAGCATTGGTCC	AGTTTATCTTGGCTTTGGTG
rs2686817.C.A	chr7:47968923-47968933	ACACTGACGACATGGTTCTACATCCCCGAAATCCATCATC	TACGGTAGCAGAGACTTGGTCTTCATGAACCCTGTTCTTAATT
rs1884214.C.T	chr14:37149132-37149142	ACACTGACGACATGGTTCTACATAAATTTTATTTCATGTTTCCTACTTG	TACGGTAGCAGAGACTTGGTCTATAGGTATTTTAAAACATTGTTTAAAACA
rs1053959.C.A	chr9:114359620-114359630	ACACTGACGACATGGTTCTACAATACAGTCCCTTCCAACTC	TACGGTAGCAGAGACTTGGTCTGTTGGCTATCCTACTAATAGTGA
rs3780190.A.G	chr9:139099069-139099079	ACACTGACGACATGGTTCTACAAGTTGTTGAAAGAATCAGCT	TACGGTAGCAGAGACTTGGTCTATGGATGTATTCATAAGACCATG
rs1019503.G.A	chr5:96254813-96254823	ACACTGACGACATGGTTCTACATACAGATAAGTTTTAGAAAGACTCAA	TACGGTAGCAGAGACTTGGTCTATTGCTCCAAAGTGGATTG
rs10883098.C.G	chr10:100218191-100218201	ACACTGACGACATGGTTCTACAGAAGCTTCCTACCTACCA	TACGGTAGCAGAGACTTGGTCTGCCTATTTGAGCTAGAACAAA
rs6979945.G.A	chr7:82386121-82386131	ACACTGACGACATGGTTCTACAAAGCAATTTGATTTATTTCTTTTGG	TACGGTAGCAGAGACTTGGTCTCTAAAGGTGAAGAATTACATGAAAAT

GenBank Target ID	Target Interval	Left Primer With Tag	Right Primer With Tag
	1		
rs147147897.CTA.C	chr16:70192457-70192469	ACACTGACGACATGGTTCTACACTCCTTCTAGACATTTCTTGC	TACGGTAGCAGAGACTTGGTCTCTAACAAGCCAACCAGAAC
rs4692500.A.C	chr4:31147268-31147278	ACACTGACGACATGGTTCTACAATGGGCTATTCCCAACTT	TACGGTAGCAGAGACTTGGTCTAGACTGGTCCATTATCTGAG
rs4727038.G.C	chr7:148979912-148979922	ACACTGACGACATGGTTCTACATGGAAAGTTTTCTACTTTGCT	TACGGTAGCAGAGACTTGGTCTGCAAAGTTCCATTAACAACTAC
rs8692.C.T	chr16:84211479-84211489	ACACTGACGACATGGTTCTACAAGCACCGAAAGCATCATA	TACGGTAGCAGAGACTTGGTCTTGTTTATCATGCCTAAGGGA
rs59508481.C.T	chr3:153839955-153839965	ACACTGACGACATGGTTCTACAAACGGGTTACTAATTACGG	TACGGTAGCAGAGACTTGGTCTCTCCTATGTACCGTCCTG
rs201131957.TTC.T	chr6:87993564-87993576	ACACTGACGACATGGTTCTACAGTGCACCAATTGTAATGAG	TACGGTAGCAGAGACTTGGTCTTAGTCTGAAGAAATCTAACAGGAT
..G.GT	chr15:52102326-52102337	ACACTGACGACATGGTTCTACAATCCTTTAATTAGTTTATGGCTTCA	TACGGTAGCAGAGACTTGGTCTCCTAAATGGGTGAAAGGG

GenBank Target ID	Target Interval	Left Primer With Tag	Right Primer With Tag
rs7096.G.A	chr9:115234607-115234617	ACACTGACGACATGGTTCTACACTGCTAGTGCTGCTTTAC	TACGGTAGCAGAGACTTGGTCTAAAGATCTAATTGGCTTTTATTAGC
..GTC.G	chr12:10311906-10311918	ACACTGACGACATGGTTCTACACCCCACTTGTCCCAAAAT	TACGGTAGCAGAGACTTGGTCTAAAGGTATCAATCCTTGGTTT
rs10486985.G.A	chr7:82386078-82386088	ACACTGACGACATGGTTCTACACACTTCACCCTTAAATAAAGC	TACGGTAGCAGAGACTTGGTCTTGCTGGTCTTACCCTAGT
rs931606.G.A	chr4:79443846-79443856	ACACTGACGACATGGTTCTACATTTCCCCTTCAAATAGGACA	TACGGTAGCAGAGACTTGGTCTAAGTCAAATTCAATTCCTCCTAG
..G.GA	chr11:108345511-108345522	ACACTGACGACATGGTTCTACAACTTGGTGCAGTCTTCTA	TACGGTAGCAGAGACTTGGTCTGAAGAACTTTGAGTCAGCC
rs30169.T.G	chr5:137190018-137190218	ACACTGACGACATGGTTCTACATTTCTGGACTCACCACTAT	TACGGTAGCAGAGACTTGGTCTCTTCAGATGTCCATTAAATTTGC
rs4870887.G.A	chr8:125061891-125061901	ACACTGACGACATGGTTCTACATTCTTTTCTTCTTTAGGGGAAG	TACGGTAGCAGAGACTTGGTCTGGTGATAGCATAACCTGG
rs113318	chr2:554	ACACTGACGACATGGTTCTACATT	TACGGTAGCAGAGACTTGGTCTG

GenBank Target ID	Target Interval	Left Primer With Tag	Right Primer With Tag
2.C.T	02221-55402231	AGGTATCTGTTCTGCTTGAT	TCTGAATAAGTATTTGATTGAAAAC
rs36004074.G.A	chr2:186665428-186665438	ACACTGACGACATGGTTCTACAAGCAGAAGAATGTCAACTTTTA	TACGGTAGCAGAGACTTGGTCTAAGGTCTGACATGTAATCTACTT
rs114154393.T.C	chr6:30999993-31000003	ACACTGACGACATGGTTCTACAGACCTGTTTCATAGGCAC	TACGGTAGCAGAGACTTGGTCTGTGAAGTCCATTCCAGGG
rs1047481.G.A	chr5:52249472-52249482	ACACTGACGACATGGTTCTACACCAGAAATGTATTTTAAAGAATTTCAG	TACGGTAGCAGAGACTTGGTCTAGTAGATTCCAATGAACCATAATG
rs2293250.G.A	chr3:33434827-33434837	ACACTGACGACATGGTTCTACAAAAACAAATGATTTGATTTACCTGA	TACGGTAGCAGAGACTTGGTCTCACCAAATTAATCAGGTTTACAG
rs10849174.T.C	chr12:5025090-5025100	ACACTGACGACATGGTTCTACAGACTGGTTGATTCTTCTGC	TACGGTAGCAGAGACTTGGTCTCCCAAACTCATAACAACAAC
rs7011928.G.A	chr8:28421981-28421991	ACACTGACGACATGGTTCTACACTCATACAAATAAGAATTGGTAGC	TACGGTAGCAGAGACTTGGTCTATCATCAGGGGCAAGTTA
rs3733690.G.C	chr5:140569523-1405695	ACACTGACGACATGGTTCTACAGAGGCATTTCTTACTAGAATCC	TACGGTAGCAGAGACTTGGTCTGAGGCAATGAATTTAAAGTCAATA

GenBank Target ID	Target Interval	Left Primer With Tag	Right Primer With Tag
	33		
rs1299331.A.G	chr3:106965488-106965498	ACACTGACGACATGGTTCTACAGGTCTTAAGACAGGAAGCTA	TACGGTAGCAGAGACTTGGTCTTCTGATTCCAGTCATAGCTC
rs2278711.C.T	chr2:47085394-47085404	ACACTGACGACATGGTTCTACATGTTCCCTTTCGGATGAG	TACGGTAGCAGAGACTTGGTCTAGTTATTTTGAGCAGATGGTT
rs11692815.G.A	chr2:68415763-68415773	ACACTGACGACATGGTTCTACAACTCGCTGTACTAAAGGAT	TACGGTAGCAGAGACTTGGTCTATGAAATTAAAAGGCTAGGAAAGA
rs116485895.T.C	chr4:69176794-69176804	ACACTGACGACATGGTTCTACACCTTCCTGATGTAATCCTCA	TACGGTAGCAGAGACTTGGTCTCAAGGAAATGGGGTTCTTT
rs2161036.G.T	chr2:186659355-186659365	ACACTGACGACATGGTTCTACAAGTGGAAGAACTTTTGAATAAGT	TACGGTAGCAGAGACTTGGTCTTTGGCAGTTTCATTAAGTGG
rs17229201.T.C	chr2:186656952-186656962	ACACTGACGACATGGTTCTACAAAATACTTAAGTATGTAGTCAAGTTAATTT	TACGGTAGCAGAGACTTGGTCTGCCTGTATCGATAAGTTTCAA
rs2230854.T.C	chr13:103275382-1032753	ACACTGACGACATGGTTCTACAGATGATGGAAACCTGCTC	TACGGTAGCAGAGACTTGGTCTACCACTTTATCACCTCATAAATCT

GenBank Target ID	Target Interval	Left Primer With Tag	Right Primer With Tag
	92		
rs4647554.A.G	chr9:97862697-97862707	ACACTGACGACATGGTTCTACAAAAAGTTCCTGGAATTTTCCA	TACGGTAGCAGAGACTTGGTCTGTTTTGAGGTATGAGCAAAGT
rs704697.T.C	chr11:33724776-33724786	ACACTGACGACATGGTTCTACAGGGGTTAAATAAGAAGTTGATGA	TACGGTAGCAGAGACTTGGTCTTCATAGGGACTTTGCCTC
rs30168.G.A	chr5:13719085-13719095	ACACTGACGACATGGTTCTACAGTTGGCAAATTTAATGGACA	TACGGTAGCAGAGACTTGGTCTATGGACATAATCATAGAAACTGA
rs55938253.A.G	chr2:186661669-186661679	ACACTGACGACATGGTTCTACAAGTTTTGATCAGACCATGAAAG	TACGGTAGCAGAGACTTGGTCTGACAAGACTTGCTTTAATGC
rs10931200.C.A	chr2:186664959-186664969	ACACTGACGACATGGTTCTACAAGAAACAAATCATTTTCTATGCAT	TACGGTAGCAGAGACTTGGTCTGAGTAGAAAATTGCCAGTCT
rs6928.C.G	chr22:22115000-22115010	ACACTGACGACATGGTTCTACACAGATACAAAGCAGTTTCAGA	TACGGTAGCAGAGACTTGGTCTCAGGTTTACCTAAAGCTTGTT
rs2282022.T.C	chr13:41374182-41374192	ACACTGACGACATGGTTCTACAACTAAACAGAATTTGGAAACCA	TACGGTAGCAGAGACTTGGTCTCCTTTGCAATCTTGAAGCTT

GenBank Target ID	Target Interval	Left Primer With Tag	Right Primer With Tag
rs1613780.G.T	chr1:144865846-144865856	ACACTGACGACATGGTTCTACATCTGGAAACATGACTCAGTT	TACGGTAGCAGAGACTTGGTCTCTAACTTCTACAGTCAGGGC
rs7893462.A.G	chr10:28228861-28228871	ACACTGACGACATGGTTCTACAAGAAAGAACGCCAACAAC	TACGGTAGCAGAGACTTGGTCTATCATTGAAAACCTTGTCAAG
rs2275145.G.A	chr14:35242824-35242834	ACACTGACGACATGGTTCTACATTTGTACCTGTCTTTTACTATGA	TACGGTAGCAGAGACTTGGTCTGGAAAACAAGGAAAATGGGA
rs9225.C.T	chr2:47086108-47086118	ACACTGACGACATGGTTCTACACCCCTCCTTAAGCATTGT	TACGGTAGCAGAGACTTGGTCTGATGCTCAAAGTATGATTTCAC
rs1130072.T.C	chr9:14615931-14615941	ACACTGACGACATGGTTCTACAAAGAAATACTTCAGTCCAACA	TACGGTAGCAGAGACTTGGTCTTTATGTATTCTGGATAACTGAAAACA
rs732774.C.T	chr13:52523804-52523814	ACACTGACGACATGGTTCTACAGGATCAATGTCAGTAGATTATTTAAAA	TACGGTAGCAGAGACTTGGTCTTGGATTGTAATCGGTTTTATCG
rs1359300.G.A	chr1:144922579-144922589	ACACTGACGACATGGTTCTACAAGTTTCCTTGAGGTTCTGA	TACGGTAGCAGAGACTTGGTCTGCTGAATTCAGATTCTTCAAG
rs254978	chr5:96230996-	ACACTGACGACATGGTTCTACAT	TACGGTAGCAGAGACTTGGTCTA

GenBank Target ID	Target Interval	Left Primer With Tag	Right Primer With Tag
2.G.T	96231006	ACCTCCTAGTGGTTTGG	CAGCGATAAGTTCCATG
rs9289122.C.G	chr3:118867043-118867053	ACACTGACGACATGGTTCTACAGGGGATTTAAATTTCAGGATG	TACGGTAGCAGAGACTTGGTCTTGAAACTTGGTAAGGATCTTC
rs4766310.G.A	chr12:5022037-5022047	ACACTGACGACATGGTTCTACAGCTACTGACCGATGTTTAAAA	TACGGTAGCAGAGACTTGGTCTGAGTCACTAGGTTTTCTGTTTT
rs4617548.A.G	chr11:16133409-16133419	ACACTGACGACATGGTTCTACACTACTTACCTGGTTTGTATGTTA	TACGGTAGCAGAGACTTGGTCTCGCTCATGATCCCAATTTT
rs10057908.T.G	chr5:102884929-102884939	ACACTGACGACATGGTTCTACAAGTAAAACTGAATTTCAAGATGC	TACGGTAGCAGAGACTTGGTCTATTTAGATACATAATAAAATTCGAGCTA
rs72901775.T.C	chr1:1647810-1647820	ACACTGACGACATGGTTCTACAAAGTCCTCAACTGACCCA	TACGGTAGCAGAGACTTGGTCTAGAAGAACAGGATAAAGCTC
rs2255546.C.T	chr5:96249111-96249121	ACACTGACGACATGGTTCTACATGGAAGGAAAGGTTATCAAGA	TACGGTAGCAGAGACTTGGTCTAGAAGATGGGTCCAATTTTC
rs2548538.T.A	chr5:96232138-96232148	ACACTGACGACATGGTTCTACACTCCCTGTTTAGGATGACTA	TACGGTAGCAGAGACTTGGTCTATTTCCTGTATTTGAGTCGG
rs177083.	chr2:324	ACACTGACGACATGGTTCTACAG	TACGGTAGCAGAGACTTGGTCTG

GenBank Target ID	Target Interval	Left Primer With Tag	Right Primer With Tag
C.T	47714-32447724	TGACCTGCTAAAGAAATACA	ATAAAGGCCAAGGAACAG
rs6876143.A.G	chr5:17380266-17380276	ACACTGACGACATGGTTCTACATGAAAACTGTCTTTCATTAGCT	TACGGTAGCAGAGACTTGGTCTAATACCATGTCTGCCTGT
rs9841174.T.C	chr3:167184874-167184884	ACACTGACGACATGGTTCTACACATGAATACAATGGCAAATAATAAT	TACGGTAGCAGAGACTTGGTCTGTGGATTTTCAAGATGCAAAG
rs116571982.G.A	chr6:31023144-31023154	ACACTGACGACATGGTTCTACAAAGACTTGAACCCTATTAGAAAA	TACGGTAGCAGAGACTTGGTCTGAATACATGTAGCTTACATGGC
rs4898.T.C	chr7:82386202-82386212	ACACTGACGACATGGTTCTACATTCTTCACCTTTAGTCAGGTTA	TACGGTAGCAGAGACTTGGTCTATCATCTAAACCAATGGCAAG
rs202026726.T.TA	chr3:86990606-86990617	ACACTGACGACATGGTTCTACAAAAGTATGTTGCATTTTAAAAGAATTAT	TACGGTAGCAGAGACTTGGTCTCACCTTTTAGAAAAGGATAAGTAAT
rs11537748.A.G	chr15:31668690-31668700	ACACTGACGACATGGTTCTACATTTTGCTGACTGTTTCCATA	TACGGTAGCAGAGACTTGGTCTGGGGTAATCCTGATTTCAAAA
rs14026.G.A	chr2:55404790-55404800	ACACTGACGACATGGTTCTACAAGTAGCTTAATGCCAACTTT	TACGGTAGCAGAGACTTGGTCTAGAAACAACCATCTTTATCTTTTG

GenBank Target ID	Target Interval	Left Primer With Tag	Right Primer With Tag
	0		
rs1780138.A.G	chr10:38501650-38501660	ACACTGACGACATGGTTCTACAAAAGTGAAGACTCCTTCTACC	TACGGTAGCAGAGACTTGGTCTAGTTAAACTATTGTTTTAAAAGCCTT
rs3800951.T.G	chr7:99758132-99758142	ACACTGACGACATGGTTCTACACATCGAAGTACTCAGCCA	TACGGTAGCAGAGACTTGGTCTTCGATTTGGGGTCTTCAT
rs1230382.T.A	chr5:96254350-96254360	ACACTGACGACATGGTTCTACATACCATTTGGTTTAAGCCTTAC	TACGGTAGCAGAGACTTGGTCTGGCAGATTCGACTTCATTT
rs201009141.CCT.C	chr7:27221060-27221072	ACACTGACGACATGGTTCTACATCTTTCCAAATGTTGCAAGA	TACGGTAGCAGAGACTTGGTCTTTTTCTTTTGCCATTAGTTGAT
rs2350917.T.C	chr8:62576972-62576982	ACACTGACGACATGGTTCTACATCTAATCTGAAATAATTAAAACCTTTCTAT	TACGGTAGCAGAGACTTGGTCTCCATAGCCATGAATAAAGGG
rs114783926.A.G	chr6:31026430-31026440	ACACTGACGACATGGTTCTACAGAGGTCACTAGCATTAGCA	TACGGTAGCAGAGACTTGGTCTGTTTGGGAGCATAGACTTT
rs36261.G.A	chr19:8327240-8327250	ACACTGACGACATGGTTCTACAGGCAACTGGACAGATCTG	TACGGTAGCAGAGACTTGGTCTCATTTGTTTTAATTGGACCCTTTT
rs108515	chr15:51	ACACTGACGACATGGTTCTACAG	TACGGTAGCAGAGACTTGGTCTT

GenBank Target ID	Target Interval	Left Primer With Tag	Right Primer With Tag
00.T.C	783816-51783826	TTACCTTTTCAATGCATCTTC	ATGATTCCAGCAATTCAGA
rs5930932.T.C	chrX:135431232-135431242	ACACTGACGACATGGTTCTACATATTCAGGTTTCCCCATCT	TACGGTAGCAGAGACTTGGTCTTTTCTAACAAGGAAGTCATCTTT
rs10226437.C.T	chr7:129984986-129984996	ACACTGACGACATGGTTCTACAAGCATAACTGCTTTCTTTTCTAT	TACGGTAGCAGAGACTTGGTCTTTCGTAGTTCAGCTGTTCA
rs7512.G.C	chr17:10583710-10583720	ACACTGACGACATGGTTCTACAGCAAAACTGAGTATGTTTACTTTA	TACGGTAGCAGAGACTTGGTCTGTGACATACCAAATGGAAATAAG
rs1168308.A.G	chr12:66742179-66742189	ACACTGACGACATGGTTCTACACCTGTTTGATAATTTGTTGCAT	TACGGTAGCAGAGACTTGGTCTATAGGTGGTGGTAAATGTTTTG
rs4260880.T.C	chr8:69020492-69020502	ACACTGACGACATGGTTCTACAAGTTTCTTATCCCAAAACATCA	TACGGTAGCAGAGACTTGGTCTCTTGCTCCGAAGATTTGTT
rs1801249.A.G	chr13:52515350-52515360	ACACTGACGACATGGTTCTACACATCGCTAGAAATGGTTAAAC	TACGGTAGCAGAGACTTGGTCTATTGGACCATTTAGAAATAACCAC
rs35382.	chr5:59821539-	ACACTGACGACATGGTTCTACAAATCCTAATAGATCATCTAATTGTT	TACGGTAGCAGAGACTTGGTCTA

GenBank Target ID	Target Interval	Left Primer With Tag	Right Primer With Tag
A.G	59821549	TATC	AGTTTATTGAGTGCAGATGTC
rs8042.T.A	chr8:28431196-28431206	ACACTGACGACATGGTTCTACAAGACAGCTAGGTCTGTCT	TACGGTAGCAGAGACTTGGTCTTGTAACAGTAAAATGGGGAAAG
rs104435 2.G.T	chr4:31147870-31147880	ACACTGACGACATGGTTCTACAACAGACCAGAAGAAATTGTC	TACGGTAGCAGAGACTTGGTCTACCTAAGATTATTTAGGTCTTTAGATT
rs174793 0.C.G	chr1:144916744-144916754	ACACTGACGACATGGTTCTACAGGTAGTAGATAACTGTTCCACT	TACGGTAGCAGAGACTTGGTCTATGGGGTAATCATTTTGACTG
rs216191 6.G.A	chr2:171073883-171073893	ACACTGACGACATGGTTCTACAAGCTCAACTCACCAGTAC	TACGGTAGCAGAGACTTGGTCTTACCAAGGCATTCAACAAG
rs221384 2.A.G	chr6:112423806-112423816	ACACTGACGACATGGTTCTACAATAGTACATGGATCATTGAGTAGA	TACGGTAGCAGAGACTTGGTCTCTGTTAGGGCAAGTCCAG
rs207020 3.G.A	chr16:70303576-70303586	ACACTGACGACATGGTTCTACAGTGCAGCATACTTACCCA	TACGGTAGCAGAGACTTGGTCTGGATTGACATGGCCTACC
rs116072 190.A.C	chr6:31023182-3102319	ACACTGACGACATGGTTCTACAAAGACTTGAACCCTATTAGAAAA	TACGGTAGCAGAGACTTGGTCTTCTTAGGAATACATGTAGCTTACA

GenBank Target ID	Target Interval	Left Primer With Tag	Right Primer With Tag
	2		
rs7562137.A.G	chr2:186625766-186625776	ACACTGACGACATGGTTCTACACTCGTTTTCTTTTAGGACATACA	TACGGTAGCAGAGACTTGGTCTCATTATCCTGTACATCTGCC
rs10141087.G.C	chr14:37147337-37147347	ACACTGACGACATGGTTCTACAGAAGATGGATTCCCTAGGT	TACGGTAGCAGAGACTTGGTCTGCCATCTGAGTTTTAAACGA
rs3453.T.C	chr21:35819059-35819069	ACACTGACGACATGGTTCTACACTCACTACTGCAAATGCAT	TACGGTAGCAGAGACTTGGTCTACCATGTACTGTTCTAAGCT
rs4349706.T.C	chr5:147516594-147516604	ACACTGACGACATGGTTCTACATAGGACGAATGACAGGAA	TACGGTAGCAGAGACTTGGTCTAAGACAAGCTCCGAAGAAT
rs546577.A.C	chr9:104172932-104172942	ACACTGACGACATGGTTCTACAAAACCTATTAATTCATCTGTTTTAACC	TACGGTAGCAGAGACTTGGTCTGCGTTTGACAGAATTGGT
rs6940018.G.C	chr6:136599389-136599399	ACACTGACGACATGGTTCTACACATAAGCTGAAAGGCCAG	TACGGTAGCAGAGACTTGGTCTCGAAAGATACATTTGAACATGAC
rs61575744.G.GT	chr1:41972271-41972282	ACACTGACGACATGGTTCTACATTTCCTTTTGTACGTCAGACT	TACGGTAGCAGAGACTTGGTCTTCTTGCAATGTTGGTATGGTA

GenBank Target ID	Target Interval	Left Primer With Tag	Right Primer With Tag
rs1958715.G.A	chr14:20528203-20528213	ACACTGACGACATGGTTCTACAATAAAGAACTTCTGGAACCTTTC	TACGGTAGCAGAGACTTGGTCTCTAGTAAAATAAACTCGGTCACT
rs1852450.C.A	chr12:16397730-16397740	ACACTGACGACATGGTTCTACATCAGTGCACTAACCAACA	TACGGTAGCAGAGACTTGGTCTTCTTTCTGATCAGCATCTTG
rs1052480.T.C	chr1:115591054-115591064	ACACTGACGACATGGTTCTACATAAAATCTATTGTTGCTTAGTGGTG	TACGGTAGCAGAGACTTGGTCTGTTCTGTTCCATGTACGATT
rs2528843.G.T	chr7:22478203-22478213	ACACTGACGACATGGTTCTACAGATTTTCTAGATGAAGCTGAAAC	TACGGTAGCAGAGACTTGGTCTCCTTCCAAGACTAGGAGTT
rs974110.A.G	chr6:66112405-66112415	ACACTGACGACATGGTTCTACACCTGCAAGGATAATCTTTCT	TACGGTAGCAGAGACTTGGTCTGTGCATCTGTTCACCAATA
rs2564990.T.C	chr5:53816778-53816788	ACACTGACGACATGGTTCTACATGGACATCAGTACTAATTTTAGTAGA	TACGGTAGCAGAGACTTGGTCTTTTAATACAGGCCAACAGC
rs35952402.A.ACTC	chr5:147516910-147516923	ACACTGACGACATGGTTCTACATGGGATTTCTTTGTCACTATCT	TACGGTAGCAGAGACTTGGTCTGTCCACCAATGGGTTTAG
rs257376.	chr7:106799993-	ACACTGACGACATGGTTCTACAG	TACGGTAGCAGAGACTTGGTCTC

GenBank Target ID	Target Interval	Left Primer With Tag	Right Primer With Tag
G.A	106800003	GACCTTGCATGGAAATTA	ATACTTTTGCTTCATGCAG
rs840385.T.C	chr5:34020354-34020364	ACACTGACGACATGGTTCTACAATGAACTTAAAACTTCAGGGG	TACGGTAGCAGAGACTTGGTCTACAGAAGTCATTTGCAAAGAA
rs316214.G.C	chr5:96497634-96497644	ACACTGACGACATGGTTCTACAATGCAAAGCAAGGGTATTTAA	TACGGTAGCAGAGACTTGGTCTCCTGTAATTATCTAATAAATTGGCTTAA

[0172] Primers were designed to be compatible with pooling, so that they could be amplified by PCR in a multiplex configuration with limited sample amounts. Massive parallel sequencing gives a high read depth, such that MAFs may be reliably detected even when the fraction of fetal material in the combination samples is 1% or lower.

[0173] 10 ml peripheral blood was collected from each of three pregnant volunteers (second trimester) into a standard EDTA blood collection tube. T-cell lymphocytes and monocytes were isolated from 5 ml of blood each using the RosetteSep™ Human T Cell Enrichment Cocktail (Stem Cell Technologies, 15061) and RosetteSep™ Human Monocyte Enrichment Cocktail (Stem Cell Technologies, 15068) according to the manufacturer's instructions. Plasma supernatant was centrifuged twice at high speed for 10 minutes to pellet and remove residual cells, and split into two samples per volunteer. 1% of resuspended monocytes, as defined by CD14 expression were added to a plasma sample to create a combination sample. By defining monocytes by CD14 expression, it is possible that the cell population also contains other phagocytes such as macrophages and neutrophils. A second plasma sample was used as a cell-

free control. DNA was purified from the combination and cell-free control samples using the QIAamp Circulating Nucleic Acid Kit (Qiagen, 55114), and from the T-cell lymphocytes using the DNeasy Blood & Tissue Kit (Qiagen, 69506). DNA concentration purified from T-cell lymphocytes was measured using the Qubit®dsDNA BR Assay Kit (Life Technologies, Q32853). After DNA extraction, library preparation was performed using the primers listed in Table 1, and massive parallel sequencing was performed on the Illumina MiSeq™ sequencer.

[0174] A summary of the procedure as performed is as follows:

[0175] Collect 20 ml blood in EDTA tube from each of 3 pregnant volunteers, second trimester.

[0176] Pipette blood into 15 or 50 ml conical tubes according to following: a) 5mL RosetteSEP Tcell; b) 10 ml RosetteSEP monocyte; and c) 5 ml RosetteSEP monocyte.

[0177] Add 10 µl 0.5M EDTA to tubes a) and c) and 20ul 0.5 M EDTA to tube b).

[0178] Add 250 µl Ab cocktail (a mixture of antibodies directed at human cell types) to tubes a) and c) and 500ul to tube b).

[0179] Incubate 10 minutes at room temperature.

[0180] Dilute with equal volume (of original sample) of SEP buffer containing density separation medium.

[0181] Layer blood over 15 ml ficoll in a SepMate tube.

[0182] Spin 10 minutes at 1200 x g.

[0183] Pour the supernatant off into three fresh 50 mL conical tubes. Set one tube aside on ice and label it 100%.

[0184] Pellet the cells in the two remaining tubes by centrifuging at 500 x g for 10 min.

[0185] Decant the supernatants into fresh 50 ml conical tube, combining the supernatants from the two remaining tubes into a single tube, label plasma.

[0186] Re-suspend the T-cell pellet in 5 ml PBS and the monocyte pellet in 10 ml phosphate-buffered saline (PBS). Set aside on ice

[0187] Spin the plasma at maximum speed for 10 min to pellet residual cells.

[0188] Decant the supernatant into a fresh 50 ml tube and repeat the spin on maximum speed for 10 min.

[0189] Decant the supernatant into a fresh 50 ml tube. Mix by inverting several times.

[0190] Divide the plasma supernatant into three 50 ml tubes with 5ml in each. Label the tubes 0%, 1%, and 5%.

[0191] Mix the monocytes well by vortexing and inverting several times. Pipette 50 µl into the 1% plasma tube and 250 µl into the 5% plasma tube.

[0192] Perform DNA extraction using a ccff kit (with carrier RNA) on the following samples: 1) 100% monocytes; 2) 5% monocytes; 3) 1% monocytes; and 4) 0% monocytes.

[0193] Perform DNA extraction using a DNeasy Blood & Tissue Kit (Qiagen, 69506) for the following samples of pelleted cells: 1) the monocyte pellet; and 2) the T-cell pellet.

[0194] Generate MAF multiplex PCR amplicons as follows: generate 48 multiplex primer pools from the MAF primers described above. Mix and dilute primers to 1 µM in pools. Modify PCR conditions as described below from the Access Array™ System for Illumina Sequencing Systems User Guide (Fluidigm).

10X Fast Start High fidelity Reaction Buffer (Roche)	0.5 µL

25 mM $MgCl_2$ (Roche)	0.9 μL
5 mM dNTPS mix	0.2 μL
5U/uL Fast Start High Fidelity Enzyme Blend (Roche	0.05 μL
PCR certified Water	1.12 μL
DNA sample	0.83 μL
Total Master mix	~4 μL
Primer mix 1 μM	1 μL

[0195] Run samples on an ABI 9700 thermocycler with a 384 well head under the thermocycle profile below.

50 °C for 2 minutes
70 °C for 20 minutes
95 °C for 10 minutes
10 cycles of the
three steps
above
95 °C for 15 seconds
60 °C for 30 seconds
72 °C for 1 minute
2 cycles of the
three steps
above
95 °C for 15 seconds
80 °C for 30 seconds
60 °C for 30 seconds

72 °C for 1 minute
8 cycles of the
four steps above
95 °C for 15 seconds
60 °C for 30 seconds
72 °C for 1 minute
2 cycles of the
three steps
above
95 °C for 15 seconds
80 °C for 30 seconds
60 °C for 30 seconds
72 °C for 1 minute
8 cycles of the
four steps above
95 °C for 15 seconds
60 °C for 30 seconds
72 °C for 1 minute
5 cycles of the
three steps
above
95 °C for 15 seconds
80 °C for 30 seconds
60 °C for 30 seconds
72 °C for 1 minute

[0196] Pool multiplex reactions from the samples into a single tube and dilute 1:100 in PCR grade water. Attach Fluidigm Barcode indexes per the Access Array™ System for Illumina Sequencing Systems User Guide (Fluidigm).

10X Fast Start High Fidelity Reaction Buffer (Roche)	2.0 µL
25 mM $MgCl_2$ (Roche)	3.6 µL
5 mM dNTP mix	1.0 µL
5 U/µL Fast Start High Fidelity Enzyme Blend (Roche)	0.2 µL
PCR certified Water	6.8 µL
Total Master mix	~14 µL
Primer mix 1 µM (Fluidigm)	4 µL
Diluted pooled PCR multiplexes	2 µL
Total	20 µL

95 ºC for 10 minutes
16 cycles of the
three steps
below
95 ºC for 15 seconds
60 ºC for 30 seconds
72 ºC for 1 minute
72 ºC for 3 minutes
4 ºC for 5 minutes

[0197] Barcode samples in triplicate and pool them at 3 x 15 μL. Purify the pools using Agencourt AmPure XP beads at a 1:1 mix and elute in 30 μL. Quantify samples with Qubit® Fluorometric Quantitiation (Life Technologies) and check on a 2100 Bioanalyzer (Agilent Technologies Genomics) for primer dimers. Normalize samples to a concentration of 10 nM for sequencing preparation.

[0198] Library purification and generation: prepare pooled, barcoded, and purified samples per the Illumina MiSeq™ manual for library generation. Dilute 10 nM samples to 2nM at 10 μL and add 10 μL 0.2 N NaOH. Incubate samples incubated at room temperature for 5 minutes and add 980 μL of HTI buffer to generate a 20 pM library. Further dilute the library to 12.5 pM in HTI buffer and load with a 1% phiX spike.

[0199] Sequencing**:** Run 600 μL of each sample on an Illumina MiSeq™ at 12.5 pM with FL1 primers prepared per the Access Array™ System for Illumina Sequencing Systems User Guide (Fluidigm). Load FL1 primers at 500 nM into an Illumina MiSeq™ reagent cartridge for sequencing and run the samples using 2x151 reads with adaptor trimming as Adapter Read 2-TGTAGAACCATGTCGT Adapter Read 1-AGACCAAGTCTCTGCT.

[0200] By a conservative estimate, the data shows 16 combination sample-specific MAFs that are informative, i.e., SNPs for which the mother was homozygous and the fetus was heterozygous (Figure 1, Table 2). This value is very near the theoretical 10% (16/167) expectation (Tynan et al., J Mol Diagn. 13(4):382-9 (2011)), confirming fetal content in the sample. Also, the presence of 16 informative MAFs falls above the level for 99.9% confidence established in Tynan et al., J Mol Diagn. 13(4):382-9 (2011). Thus, the fetal DNA content in the combination sample was confirmed. Figures 1 and 2 show potentially informative alleles from two assays plotted against the percent of reads of each allele in maternal and combination samples. A rough estimate of informative MAFs can be seen by the boxed off sections at the top and bottom of the graph. In general, informative MAFs are seen where an allele in a combination sample shows major and minor frequencies. The significance of these differenced can be determined by a number of statistical analyses

known in the art (for example, values that are five standard deviations from the mean value or other statistical measures). Figure 2 in particular shows tight clustering of data around the 50% mark, demonstrating a low standard deviation. Moreover, Figure 5 shows the improvement in data between a combination sample (cell-free bodily fluid plus 1% monocytes) when compared to a cell-free bodily fluid or T-cell sample alone (Figures 3 and 4, respectively). Figures 6 and 7 are simplified to show only the informative markers, and demonstrate that the combination sample produces "cleaner" data than a cell-free bodily fluid sample without monocytes. Figures 8 and 9 show exemplary informative markers. Both depict a maternal only sample wherein the reference allele reads are 100% and the alternative allele reads are 100%, and a combination sample that contains reads of the alterative allele (and, necessarily, fewer reads of the reference allele such that the values add to 100%). By contrast, the exemplary non-informative marker depicted in Figure 10 has reads of approximately 50% of reference and alternative alleles in both the maternal only and the combination sample. In addition, reproducibility in combination samples is better than in corresponding cell-free bodily fluid samples (Table 3). In Table 3, the differences between Assay 1 and Assay 2 may be because a low-performing marker was removed in Assay 2 and one of the markers in Assay 2 also was rejected due to a threshold effect with the computational percentage (to include all base calls at a position rather than only the reference and alternative markers).

[0201] It has been proposed that a qualitative size difference exists between cell-free circulating DNA and maternal DNA (Fan et al., Clin Chem. 56(8):1279-86 (2010)). The nature of this size difference may result from the pathway by which the DNA is produced. Fetal circulating cell-free DNA is thought to be almost exclusively generated via apoptosis of liberated fetal cells, and it exists in the plasma at nucleosomal lengths preserved in apoptotic bodies. By contrast, maternal circulating cell-free DNA is more likely to be a mixture of cells produced from apoptosis and necrosis as well as autophagy and mitotic catastrophe. The DNA produced during the highly ordered apoptotic process may exist in a less degraded state than the DNA produced during other cell death pathways. DNA produced during non-apoptitic cell death is exposed to a variety of harmful subcellular components including DNA damaging reactive oxygen species. Without wishing to be bound by theory, the

increased relative degradation of the maternal circulating cell-free DNA may contribute to noise in amplification-based assays aimed at detecting aneuploidies from circulating cell-free DNA. Thus, the methods of using combinations of two or more different components selected from cell-free bodily fluids, phagocytic cells, circulating vesicles, and circulating diseased cells may contribute higher quality maternal DNA and reduce signal variance.

Table 2: 16 informative markers

Sample	**Chromosome**	**Position**	**Ref. allele**	**Alt. allele**	**A counts**	**C counts**	**G counts**	**T counts**
maternal	chr3	113804978	G	A	187335	10	464	23
combination	chr3	113804978	G	A	126447	6	10980	29
maternal	chr6	42627433	G	A	223746	111	704	144
combination	chr6	42627433	G	A	262863	294	13527	292
maternal	chr3	42787468	A	G	111491	15	401	21
combination	chr3	42787468	A	G	269893	39	14459	327
maternal	chr5	13701535	T	C	4	83731	0	35
combination	chr5	13701535	T	C	0	31414	0	2162
maternal	chr9	115234610	G	A	172747	24	655	98
combination	chr9	115234610	G	A	116353	40	3831	192
maternal	chr5	13719021	T	G	100	4	166680	7
combination	chr5	13719021	T	G	240	65	311800	11697

maternal	chr8	125061894	G	A	19879	16	103	4
combination	chr8	125061894	G	A	45848	21	4760	25
maternal	chr4	69176797	T	C	67	599	16	130431
combination	chr4	69176797	T	C	93	3729	40	120555
maternal	chr5	13719088	G	A	280520	36	1267	40
combination	chr5	13719088	G	A	226691	50	8320	130
maternal	chr13	41374185	T	C	0	82	2	38046
combination	chr13	41374185	T	C	18	4788	16	66696
maternal	chr2	32447717	C	T	99	1212	16	280044
combination	chr2	32447717	C	T	143	21563	15	217092
maternal	chr5	17380269	A	G	18	0	12863	1
combination	chr5	17380269	A	G	1725	0	14276	7
maternal	chr7	99758135	T	G	278	19	313255	28
combination	chr7	99758135	T	G	429	114	455545	20924
maternal	chr17	10583713	G	C	306	13	418048	10
combination	chr17	10583713	G	C	277	11325	310119	20
maternal	chr12	66742182	A	G	20	4	44686	2
combination	chr12	66742182	A	G	998	0	26251	2

Ref. Allele is the references allele, Alt. allele is the alternate allele, and A, C, G, and T counts represent the number of reads of each of the four DNA bases in the samples

Table 3: Comparison of results, four samples

	Assay 1		Assay 2	
	T-cell vs. cell-free bodily fluid	T-cell vs. combination sample (cell free + 1% monocyte)	T-cell vs. cell-free bodily fluid	T-cell vs. combination sample (cell free + 1% monocyte)
# Informative markers	14	8	38	33
Mean estimate of fetal DNA content	5.8%	5.6%	10.7%	6.05%
Standard deviation of fetal DNA estimate	2.54%	2.07%	4.3%	1.75%

[0202] The proof-of-principle here shows that diagnosis of, for example, fetal aneuploidy can be accomplished by selecting an appropriate panel of informative MAF SNPs weighted to the chromosomes of interest (e.g., chromosome 21), determination of fetal content by read depth computation, and comparison to control standards.

What is Claimed is:

1. A method for diagnosing or aiding in the diagnosis of a disease or condition in a subject comprising:

a) determining a first profile of one or more markers of the disease or condition from a sample comprising two or more different components selected from the group consisting of: a cell-free bodily fluid isolated from the subject, a population of phagocytic cells isolated from the subject, a population of phagocytic cells having a DNA content more than 2n (>2n phagocytic cells) isolated from the subject, a population of circulating vesicles isolated from the subject, and a population of circulating diseased cells isolated from the subject;

b) determining a second profile of at least one of the one or more markers from a control comprising a component selected from the group consisting of: a population of phagocytic cells having a DNA content of 2n (=2n phagocytic cells) isolated from the subject, a population of non-phagocytic cells isolated from the subject, and a population of control cells isolated from the subject, wherein the control cells are substantially free of cells affected by the disease or condition; and

c) identifying a difference between the first and second profiles, wherein the difference is indicative of the presence of said disease or condition in the subject.

2. A method for assessing the risk of developing a disease or condition in a subject comprising:

a) determining a first profile of one or more markers of the disease or condition from a sample comprising two or more different components selected from the group consisting of: a cell-free bodily fluid isolated from the subject, a population of phagocytic cells isolated from the subject, a population of phagocytic cells having a DNA content more than 2n (>2n phagocytic cells) isolated from the subject, a population of circulating vesicles isolated from the subject, and a population of circulating diseased cells isolated from the subject;

b) determining a second profile of at least one of the one or more markers from a control comprising a component selected from the group consisting of: a population of phagocytic cells having a DNA content of 2n (=2n phagocytic cells) isolated from the subject, a population of non-phagocytic cells isolated from the

subject, and a population of control cells isolated from the subject, wherein the control cells are substantially free of cells affected by the disease or condition; and

c) identifying a difference between the first and second profiles, wherein the difference is indicative of the risk of developing said disease or condition in the subject.

3. A method for prognosing or aiding in the prognosis of a disease or condition in a subject comprising:

a) determining a first profile of one or more markers of the disease or condition from a sample comprising two or more different components selected from the group consisting of: a cell-free bodily fluid isolated from the subject, a population of phagocytic cells isolated from the subject, a population of phagocytic cells having a DNA content more than 2n (>2n phagocytic cells) isolated from the subject, a population of circulating vesicles isolated from the subject, and a population of circulating diseased cells isolated from the subject;

b) determining a second profile of at least one of the one or more markers from a control comprising a component selected from the group consisting of: a population of phagocytic cells having a DNA content of 2n (=2n phagocytic cells) isolated from the subject, a population of non-phagocytic cells isolated from the subject, and a population of control cells isolated from the subject, wherein the control cells are substantially free of cells affected by the disease or condition; and

c) identifying a difference between the first and second profiles, wherein the difference is indicative of the prognosis of said disease or condition in the subject.

4. A method for assessing the efficacy of a treatment for a disease or condition in a subject comprising:

a) determining a first profile of one or more markers of the disease or condition from a first sample comprising two or more different components selected from the group consisting of: a cell-free bodily fluid isolated from the subject before the treatment, a population of phagocytic cells isolated from the subject before the treatment, a population of phagocytic cells having a DNA content more than 2n (>2n phagocytic cells) isolated from the subject before the treatment, a population of circulating vesicles isolated from the subject before the treatment, and a population of circulating diseased cells isolated from the subject before the treatment;

determining a second profile of at least one of the one or more markers from a first control comprising a component selected from the group consisting of: a population of phagocytic cells having a DNA content of 2n (=2n phagocytic cells) isolated from the subject before the treatment, a population of non-phagocytic cells isolated from the subject before the treatment, and a population of control cells isolated from the subject, wherein the control cells are substantially free of cells affected by the disease or condition, before the treatment;

identifying a difference between the first and second profiles;

b) determining a third profile of one or more markers of the disease or condition from a second sample comprising two or more different components selected from the group consisting of: a cell-free bodily fluid isolated from the subject after the treatment, a population of phagocytic cells isolated from the subject after the treatment, a population of phagocytic cells having a DNA content more than 2n (>2n phagocytic cells) isolated from the subject after the treatment, a population of circulating vesicles isolated from the subject after the treatment, and a population of circulating diseased cells isolated from the subject after the treatment;

determining a fourth profile of at least one of the one or more markers from a second control comprising a component selected from the group consisting of: a population of phagocytic cells having a DNA content of 2n (=2n phagocytic cells) isolated from the subject after the treatment, a population of non-phagocytic cells isolated from the subject after the treatment, and a population of control cells isolated from the subject, wherein the control cells are substantially free of cells affected by the disease or condition, after the treatment;

identifying a difference between the third and fourth profiles; and

c) identifying a difference between the difference identified in a) and the difference identified in b) wherein the identified difference in c) is indicative of the efficacy of the treatment for said disease or condition in the subject.

5. A method for monitoring the progression or regression of a disease or condition in a subject comprising:

a) determining a first profile of one or more markers of the disease or condition from a first sample comprising two or more different components selected from the group consisting of: a cell-free bodily fluid isolated from the subject at a first

time point, a population of phagocytic cells isolated from the subject at a first time point, a population of phagocytic cells having a DNA content more than 2n (>2n phagocytic cells) isolated from the subject at a first time point, a population of circulating vesicles isolated from the subject at a first time point, and a population of circulating diseased cells isolated from the subject at a first time point;

determining a second profile of at least one of the one or more markers from a first control comprising a component selected from the group consisting of: a population of phagocytic cells having a DNA content of 2n (=2n phagocytic cells) isolated from the subject at a first time point, a population of non-phagocytic cells isolated from the subject at a first time point, and a population of control cells isolated from the subject, wherein the control cells are substantially free of cells affected by the disease or condition, at a first time point;

identifying a difference between the first and second profiles;

b) determining a third profile of one or more markers of the disease or condition from a second sample comprising two or more different components selected from the group consisting of: a cell-free bodily fluid isolated from the subject at a second time point, a population of phagocytic cells isolated from the subject at a second time point, a population of phagocytic cells having a DNA content more than 2n (>2n phagocytic cells) isolated from the subject at a second time point, a population of circulating vesicles isolated from the subject at a second time point, and a population of circulating diseased cells isolated from the subject at a second time point;

determining a fourth profile of at least one of the one or more markers from a second control comprising a component selected from the group consisting of: a population of phagocytic cells having a DNA content of 2n (=2n phagocytic cells) isolated from the subject at a second time point, a population of non-phagocytic cells isolated from the subject at a second time point, and a population of control cells isolated from the subject, wherein the control cells are substantially free of cells affected by the disease or condition, at a second time point;

identifying a difference between the third and fourth profiles; and

c) identifying a difference between the difference identified in a) and the difference identified in b) wherein the identified difference in c) is indicative of the progression or regression of said disease or condition in the subject.

6. A method for identifying a compound capable of ameliorating or treating a disease or condition in a subject comprising:

a) determining a first profile of one or more markers of the disease or condition from a first sample comprising two or more different components selected from the group consisting of: a cell-free bodily fluid isolated from the subject before administering the compound to the subject, a population of phagocytic cells isolated from the subject before administering the compound to the subject, a population of phagocytic cells having a DNA content more than 2n (>2n phagocytic cells) isolated from the subject before administering the compound to the subject, a population of circulating vesicles isolated from the subject before administering the compound to the subject, and a population of circulating diseased cells isolated from the subject before administering the compound to the subject;

determining a second profile of at least one of the one or more markers from a first control comprising a component selected from the group consisting of: a population of phagocytic cells having a DNA content of 2n (=2n phagocytic cells) isolated from the subject before administering the compound to the subject, a population of non-phagocytic cells isolated from the subject before administering the compound to the subject, and a population of control cells isolated from the subject, wherein the control cells are substantially free of cells affected by the disease or condition, before administering the compound to the subject;

identifying a difference between the first and second profiles;

b) determining a third profile of one or more markers of the disease or condition from a second sample comprising two or more different components selected from the group consisting of: a cell-free bodily fluid isolated from the subject after administering the compound to the subject, a population of phagocytic cells isolated from the subject after administering the compound to the subject, a population of phagocytic cells having a DNA content more than 2n (>2n phagocytic cells) isolated from the subject after administering the compound to the subject, a population of circulating vesicles isolated from the subject after administering the

compound to the subject, and a population of circulating diseased cells isolated from the subject after administering the compound to the subject;

determining a fourth profile of at least one of the one or more markers from a second control comprising a component selected from the group consisting of: a population of phagocytic cells having a DNA content of 2n (=2n phagocytic cells) isolated from the subject after administering the compound to the subject, a population of non-phagocytic cells isolated from the subject after administering the compound to the subject, and a population of control cells isolated from the subject, wherein the control cells are substantially free of cells affected by the disease or condition, after administering the compound to the subject;

identifying a difference between the third and fourth profiles; and

c) identifying a difference between the difference identified in a) and the difference identified in b) wherein the identified difference in c) indicates that the compound is capable of ameliorating or treating said disease or condition in the subject.

7. A method for assessing the efficacy of a treatment for a disease or condition in a subject comprising:

a) determining a first profile of one or more markers of the disease or condition from a first sample comprising two or more different components selected from the group consisting of: a cell-free bodily fluid isolated from the subject before the treatment, a population of phagocytic cells isolated from the subject before the treatment, a population of phagocytic cells having a DNA content more than 2n (>2n phagocytic cells) isolated from the subject before the treatment, a population of circulating vesicles isolated from the subject before the treatment, and a population of circulating diseased cells isolated from the subject before the treatment;

b) determining a second profile of one or more markers of the disease or condition from a second sample comprising two or more different components selected from the group consisting of: a cell-free bodily fluid isolated from the subject after the treatment, a population of phagocytic cells isolated from the subject after the treatment, a population of phagocytic cells having a DNA content more than 2n (>2n phagocytic cells) isolated from the subject after the treatment, a population of

circulating vesicles isolated from the subject after the treatment, and a population of circulating diseased cells isolated from the subject; and

c) identifying a difference between the first profile and the second profile, wherein the identified difference is indicative of the efficacy of the treatment for said disease or condition in the subject.

8. A method for monitoring the progression or regression of a disease or condition in a subject comprising:

a) determining a first profile of one or more markers of the disease or condition from a first sample comprising two or more different components selected from the group consisting of: a cell-free bodily fluid isolated from the subject at a first time point, a population of phagocytic cells isolated from the subject at a first time point, a population of phagocytic cells having a DNA content more than 2n (>2n phagocytic cells) isolated from the subject at a first time point, a population of circulating vesicles isolated from the subject at a first time point, and a population of circulating diseased cells isolated from the subject at a first time point;

b) determining a second profile of one or more markers of the disease or condition from a second sample comprising two or more different components selected from the group consisting of: a cell-free bodily fluid isolated from the subject at a second time point, a population of phagocytic cells isolated from the subject at a second time point, a population of phagocytic cells having a DNA content more than 2n (>2n phagocytic cells) isolated from the subject at a second time point, a population of circulating vesicles isolated from the subject at a second time point, and a population of circulating diseased cells isolated from the subject at a second time point; and

c) identifying a difference between the first profile and the second profile, wherein the identified difference is indicative of the progression or regression of said disease or condition in the subject.

9. A method for identifying a compound capable of ameliorating or treating a disease or condition in a subject comprising:

a) determining a first profile of one or more markers of the disease or condition from a first sample comprising two or more different components selected from the group consisting of: a cell-free bodily fluid isolated from the subject before

administering the compound to the subject, a population of phagocytic cells isolated from the subject before administering the compound to the subject, a population of phagocytic cells having a DNA content more than 2n (>2n phagocytic cells) isolated from the subject before administering the compound to the subject, a population of circulating vesicles isolated from the subject before administering the compound to the subject, and a population of circulating diseased cells isolated from the subject before administering the compound to the subject;

b) determining a second profile of one or more markers of the disease or condition from a second sample comprising two or more different components selected from the group consisting of: a cell-free bodily fluid isolated from the subject after administering the compound to the subject, a population of phagocytic cells isolated from the subject after administering the compound to the subject, a population of phagocytic cells having a DNA content more than 2n (>2n phagocytic cells) isolated from the subject after administering the compound to the subject, a population of circulating vesicles isolated from the subject after administering the compound to the subject, and a population of circulating diseased cells isolated from the subject after administering the compound to the subject; and

c) identifying a difference between the first profile and the second profile, wherein the identified difference indicates that the compound is capable of ameliorating or treating said disease or condition in the subject.

10. A method for diagnosing or aiding in the diagnosis of a disease or condition in a subject comprising:

a) determining a first profile of one or more markers of the disease or condition from a sample comprising components selected from the group consisting of: an analyte isolated from a cell-free bodily fluid isolated from the subject, an analyte isolated from a population of phagocytic cells isolated from the subject, an analyte isolated from a population of phagocytic cells having a DNA content more than 2n (>2n phagocytic cells) isolated from the subject, an analyte isolated from a population of circulating vesicles isolated from the subject, and an analyte isolated from a population of circulating diseased cells isolated from the subject;

b) determining a second profile of at least one of the one or more markers from a control comprising a component selected from the group consisting of: an

analyte isolated from a population of phagocytic cells having a DNA content of 2n (=2n phagocytic cells) isolated from the subject, an analyte isolated from a population of non-phagocytic cells isolated from the subject, and an analyte isolated from a population of control cells isolated from the subject, wherein the control cells are substantially free of cells affected by the disease or condition; and

c) identifying a difference between the first and second profiles, wherein the difference is indicative of the presence of said disease or condition in the subject.

11. A method for assessing the risk of developing a disease or condition in a subject comprising:

a) determining a first profile of one or more markers of the disease or condition from a sample comprising two or more different components selected from the group consisting of: an analyte isolated from a cell-free bodily fluid isolated from the subject, an analyte isolated from a population of phagocytic cells isolated from the subject, an analyte isolated from a population of phagocytic cells having a DNA content more than 2n (>2n phagocytic cells) isolated from the subject, an analyte isolated from a population of circulating vesicles isolated from the subject, and an analyte isolated from a population of circulating diseased cells isolated from the subject;

b) determining a second profile of at least one of the one or more markers from a control comprising a component selected from the group consisting of: an analyte isolated from a population of phagocytic cells having a DNA content of 2n (=2n phagocytic cells) isolated from the subject, an analyte isolated from a population of non-phagocytic cells isolated from the subject, and an analyte isolated from a population of control cells isolated from the subject, wherein the control cells are substantially free of cells affected by the disease or condition; and

c) identifying a difference between the first and second profiles, wherein the difference is indicative of the risk of developing said disease or condition in the subject.

12. A method for prognosing or aiding in the prognosis of a disease or condition in a subject comprising:

a) determining a first profile of one or more markers of the disease or condition from a sample comprising two or more different components selected from

the group consisting of: an analyte isolated from a cell-free bodily fluid isolated from the subject, an analyte isolated from a population of phagocytic cells isolated from the subject, an analyte isolated from a population of phagocytic cells having a DNA content more than 2n (>2n phagocytic cells) isolated from the subject, an analyte isolated from a population of circulating vesicles isolated from the subject, and an analyte isolated from a population of circulating diseased cells isolated from the subject;

b) determining a second profile of at least one of the one or more markers from a control comprising a component selected from the group consisting of: an analyte isolated from a population of phagocytic cells having a DNA content of 2n (=2n phagocytic cells) isolated from the subject, an analyte isolated from a population of non-phagocytic cells isolated from the subject, and an analyte isolated from a population of control cells isolated from the subject, wherein the control cells are substantially free of cells affected by the disease or condition; and

c) identifying a difference between the first and second profiles, wherein the difference is indicative of the prognosis of said disease or condition in the subject.

13. A method for assessing the efficacy of a treatment for a disease or condition in a subject comprising:

a) determining a first profile of one or more markers of the disease or condition from a first sample comprising two or more different components selected from the group consisting of: an analyte isolated from a cell-free bodily fluid isolated from the subject before the treatment, an analyte isolated from a population of phagocytic cells isolated from the subject before the treatment, an analyte isolated from a population of phagocytic cells having a DNA content more than 2n (>2n phagocytic cells) isolated from the subject before the treatment, an analyte isolated from a population of circulating vesicles isolated from the subject before the treatment, and an analyte isolated from a population of circulating diseased cells isolated from the subject before the treatment;

determining a second profile of at least one of the one or more markers from a first control comprising a component selected from the group consisting of: an analyte isolated from a population of phagocytic cells having a DNA content of 2n (=2n phagocytic cells) isolated from the subject before the treatment, an analyte

isolated from a population of non-phagocytic cells isolated from the subject before the treatment, and an analyte isolated from a population of control cells isolated from the subject, wherein the control cells are substantially free of cells affected by the disease or condition, before the treatment;

identifying a difference between the first and second profiles;

b) determining a third profile of one or more markers of the disease or condition from a second sample comprising two or more different components selected from the group consisting of: an analyte isolated from a cell-free bodily fluid isolated from the subject after the treatment, an analyte isolated from a population of phagocytic cells isolated from the subject after the treatment, an analyte isolated from a population of phagocytic cells having a DNA content more than 2n (>2n phagocytic cells) isolated from the subject after the treatment, an analyte isolated from a population of circulating vesicles isolated from the subject after the treatment, and an analyte isolated from a population of circulating diseased cells isolated from the subject after the treatment;

determining a fourth profile of at least one of the one or more markers from a second control comprising a component selected from the group consisting of: an analyte isolated from a population of phagocytic cells having a DNA content of 2n (=2n phagocytic cells) isolated from the subject after the treatment, an analyte isolated from a population of non-phagocytic cells isolated from the subject after the treatment, and an analyte isolated from a population of control cells isolated from the subject, wherein the control cells are substantially free of cells affected by the disease or condition, after the treatment;

identifying a difference between the third and fourth profiles; and

c) identifying a difference between the difference identified in a) and the difference identified in b) wherein the identified difference in c) is indicative of the efficacy of the treatment for said disease or condition in the subject.

14. A method for monitoring the progression or regression of a disease or condition in a subject comprising:

a) determining a first profile of one or more markers of the disease or condition from a first sample comprising two or more different components selected from the group consisting of: an analyte isolated from a cell-free bodily fluid isolated

from the subject at a first time point, an analyte isolated from a population of phagocytic cells isolated from the subject at a first time point, an analyte isolated from a population of phagocytic cells having a DNA content more than 2n (>2n phagocytic cells) isolated from the subject at a first time point, an analyte isolated from a population of circulating vesicles isolated from the subject at a first time point, and an analyte isolated from a population of circulating diseased cells isolated from the subject at a first time point;

determining a second profile of at least one of the one or more markers from a first control comprising a component selected from the group consisting of: an analyte isolated from a population of phagocytic cells having a DNA content of 2n (=2n phagocytic cells) isolated from the subject at a first time point, an analyte isolated from a population of non-phagocytic cells isolated from the subject at a first time point, and an analyte isolated from a population of control cells isolated from the subject, wherein the control cells are substantially free of cells affected by the disease or condition, at a first time point;

identifying a difference between the first and second profiles;

b) determining a third profile of one or more markers of the disease or condition from a second sample comprising two or more different components selected from the group consisting of: an analyte isolated from a cell-free bodily fluid isolated from the subject at a second time point, an analyte isolated from a population of phagocytic cells isolated from the subject at a second time point, an analyte isolated from a population of phagocytic cells having a DNA content more than 2n (>2n phagocytic cells) isolated from the subject at a second time point, an analyte isolated from a population of circulating vesicles isolated from the subject at a second time point, and an analyte isolated from a population of circulating diseased cells isolated from the subject at a second time point;

determining a fourth profile of at least one of the one or more markers from a second control comprising a component selected from the group consisting of: an analyte isolated from a population of phagocytic cells having a DNA content of 2n (=2n phagocytic cells) isolated from the subject at a second time point, an analyte isolated from a population of non-phagocytic cells isolated from the subject at a second time point, and an analyte isolated from a population of control cells isolated

from the subject, wherein the control cells are substantially free of cells affected by the disease or condition, at a second time point;

identifying a difference between the third and fourth profiles; and

c) identifying a difference between the difference identified in a) and the difference identified in b) wherein the identified difference in c) is indicative of the progression or regression of said disease or condition in the subject.

15. A method for identifying a compound capable of ameliorating or treating a disease or condition in a subject comprising:

a) determining a first profile of one or more markers of the disease or condition from a first sample comprising two or more different components selected from the group consisting of: an analyte isolated from a cell-free bodily fluid isolated from the subject before administering the compound to the subject, an analyte isolated from a population of phagocytic cells isolated from the subject before administering the compound to the subject, an analyte isolated from a population of phagocytic cells having a DNA content more than 2n (>2n phagocytic cells) isolated from the subject before administering the compound to the subject, an analyte isolated from a population of circulating vesicles isolated from the subject before administering the compound to the subject, and an analyte isolated from a population of circulating diseased cells isolated from the subject before administering the compound to the subject;

determining a second profile of at least one of the one or more markers from a first control comprising a component selected from the group consisting of: an analyte isolated from a population of phagocytic cells having a DNA content of 2n (=2n phagocytic cells) isolated from the subject before administering the compound to the subject, an analyte isolated from a population of non-phagocytic cells isolated from the subject before administering the compound to the subject, and an analyte isolated from a population of control cells isolated from the subject, wherein the control cells are substantially free of cells affected by the disease or condition, before administering the compound to the subject;

identifying a difference between the first and second profiles;

b) determining a third profile of one or more markers of the disease or condition from a second sample comprising two or more different components

selected from the group consisting of: an analyte isolated from a cell-free bodily fluid isolated from the subject after administering the compound to the subject, an analyte isolated from a population of phagocytic cells isolated from the subject after administering the compound to the subject, an analyte isolated from a population of phagocytic cells having a DNA content more than 2n (>2n phagocytic cells) isolated from the subject after administering the compound to the subject, an analyte isolated from a population of circulating vesicles isolated from the subject after administering the compound to the subject, and an analyte isolated from a population of circulating diseased cells isolated from the subject after administering the compound to the subject;

determining a fourth profile of at least one of the one or more markers from a second control comprising a component selected from the group consisting of: an analyte isolated from a population of phagocytic cells having a DNA content of 2n (=2n phagocytic cells) isolated from the subject after administering the compound to the subject, an analyte isolated from a population of non-phagocytic cells isolated from the subject after administering the compound to the subject, and an analyte isolated from a population of control cells isolated from the subject, wherein the control cells are substantially free of cells affected by the disease or condition, after administering the compound to the subject;

identifying a difference between the third and fourth profiles; and

c) identifying a difference between the difference identified in a) and the difference identified in b) wherein the identified difference in c) indicates that the compound is capable of ameliorating or treating said disease or condition in the subject.

16. A method for assessing the efficacy of a treatment for a disease or condition in a subject comprising:

a) determining a first profile of one or more markers of the disease or condition from a first sample comprising two or more different components selected from the group consisting of: an analyte isolated from a cell-free bodily fluid isolated from the subject before the treatment, a population of phagocytic cells isolated from the subject before the treatment, an analyte isolated from a population of phagocytic cells having a DNA content more than 2n (>2n phagocytic cells) isolated from the

subject before the treatment, an analyte isolated from a population of circulating vesicles isolated from the subject before the treatment, and an analyte isolated from a population of circulating diseased cells isolated from the subject before the treatment;

b) determining a second profile of one or more markers of the disease or condition from a second sample comprising two or more different components selected from the group consisting of: an analyte isolated from a cell-free bodily fluid isolated from the subject after the treatment, an analyte isolated from a population of phagocytic cells isolated from the subject after the treatment, an analyte isolated from a population of phagocytic cells having a DNA content more than 2n (>2n phagocytic cells) isolated from the subject after the treatment, an analyte isolated from a population of circulating vesicles isolated from the subject after the treatment, and an analyte isolated from a population of control cells isolated from the subject, wherein the control cells are substantially free of cells affected by the disease or condition; and

c) identifying a difference between the first profile and the second profile, wherein the identified difference is indicative of the efficacy of the treatment for said disease or condition in the subject.

17. A method for monitoring the progression or regression of a disease or condition in a subject comprising:

a) determining a first profile of one or more markers of the disease or condition from a first sample comprising two or more different components selected from the group consisting of: an analyte isolated from a cell-free bodily fluid isolated from the subject at a first time point, an analyte isolated from a population of phagocytic cells isolated from the subject at a first time point, an analyte isolated from a population of phagocytic cells having a DNA content more than 2n (>2n phagocytic cells) isolated from the subject at a first time point, an analyte isolated from a population of circulating vesicles isolated from the subject at a first time point, and an analyte isolated from a population of circulating diseased cells isolated from the subject at a first time point;

b) determining a second profile of one or more markers of the disease or condition from a second sample comprising two or more different components selected from the group consisting of: an analyte isolated from a cell-free bodily fluid isolated from the subject at a second time point, an analyte isolated from a population

of phagocytic cells isolated from the subject at a second time point, an analyte isolated from a population of phagocytic cells having a DNA content more than 2n (>2n phagocytic cells) isolated from the subject at a second time point, an analyte isolated from a population of circulating vesicles isolated from the subject at a second time point, and an analyte isolated from a population of circulating diseased cells isolated from the subject at a second time point; and

c) identifying a difference between the first profile and the second profile, wherein the identified difference is indicative of the progression or regression of said disease or condition in the subject.

18. A method for identifying a compound capable of ameliorating or treating a disease or condition in a subject comprising:

a) determining a first profile of one or more markers of the disease or condition from a first sample comprising two or more different components selected from the group consisting of: an analyte isolated from a cell-free bodily fluid isolated from the subject before administering the compound to the subject, an analyte isolated from a population of phagocytic cells isolated from the subject before administering the compound to the subject, an analyte isolated from a population of phagocytic cells having a DNA content more than 2n (>2n phagocytic cells) isolated from the subject before administering the compound to the subject, an analyte isolated from a population of circulating vesicles isolated from the subject before administering the compound to the subject, and an analyte isolated from a population of circulating diseased cells isolated from the subject before administering the compound to the subject;

b) determining a second profile of one or more markers of the disease or condition from a second sample comprising two or more different components selected from the group consisting of: an analyte isolated from a cell-free bodily fluid isolated from the subject after administering the compound to the subject, an analyte isolated from a population of phagocytic cells isolated from the subject after administering the compound to the subject, an analyte isolated from a population of phagocytic cells having a DNA content more than 2n (>2n phagocytic cells) isolated from the subject after administering the compound to the subject, an analyte isolated from a population of circulating vesicles isolated from the subject after administering

the compound to the subject, and an analyte isolated from a population of circulating diseased cells isolated from the subject after administering the compound to the subject; and

c) identifying a difference between the first profile and the second profile, wherein the identified difference indicates that the compound is capable of ameliorating or treating said disease or condition in the subject.

19. A method for diagnosing or aiding in the diagnosis of a disease or condition in a subject comprising:

a) determining a first profile of one or more markers of the disease or condition from a sample comprising two or more different components selected from the group consisting of: a cell-free bodily fluid isolated from the subject, a population of phagocytic cells isolated from the subject, a population of phagocytic cells having a DNA content more than 2n (>2n phagocytic cells) isolated from the subject, a population of circulating vesicles isolated from the subject, and a population of circulating diseased cells isolated from the subject; and

b) identifying a difference between the first profile and a second profile of at least one of the one or more markers from a repository of said markers of said disease or condition, wherein the difference is indicative of the presence of said disease or condition in the subject.

20. A method for assessing the risk of developing a disease or condition in a subject comprising:

a) determining a first profile of one or more markers of the disease or condition from a sample comprising two or more different components selected from the group consisting of: a cell-free bodily fluid isolated from the subject, a population of phagocytic cells isolated from the subject, a population of phagocytic cells having a DNA content more than 2n (>2n phagocytic cells) isolated from the subject, a population of circulating vesicles isolated from the subject, and a population of circulating diseased cells isolated from the subject; and

b) identifying a difference between the first profile and a second profile of at least one of the one or more markers from a repository of said markers of said disease or condition , wherein the difference is indicative of the risk of developing said disease or condition in the subject.

21. A method for prognosing or aiding in the prognosis of a disease or condition in a subject comprising:

a) determining a first profile of one or more markers of the disease or condition from a sample comprising two or more different components selected from the group consisting of: a cell-free bodily fluid isolated from the subject, a population of phagocytic cells isolated from the subject, a population of phagocytic cells having a DNA content more than 2n (>2n phagocytic cells) isolated from the subject, a population of circulating vesicles isolated from the subject, and a population of circulating diseased cells isolated from the subject; and

b) identifying a difference between the first profile and a second profile of at least one of the one or more markers from a repository of said markers of said disease or condition, wherein the difference is indicative of the prognosis of said disease or condition in the subject.

22. A method for diagnosing or aiding in the diagnosis of a disease or condition in a subject comprising:

a) determining a first profile of one or more markers of the disease or condition from a sample comprising components selected from the group consisting of: an analyte isolated from a cell-free bodily fluid isolated from the subject, an analyte isolated from a population of phagocytic cells isolated from the subject, an analyte isolated from a population of phagocytic cells having a DNA content more than 2n (>2n phagocytic cells) isolated from the subject, an analyte isolated from a population of circulating vesicles isolated from the subject, and an analyte isolated from a population of circulating diseased cells isolated from the subject; and

b) identifying a difference between the first profile and a second profile of at least one of the one or more markers from a repository of said markers of said disease or condition , wherein the difference is indicative of the presence of said disease or condition in the subject.

23. A method for assessing the risk of developing a disease or condition in a subject comprising:

a) determining a first profile of one or more markers of the disease or condition from a sample comprising two or more different components selected from the group consisting of: an analyte isolated from a cell-free bodily fluid isolated from

the subject, an analyte isolated from a population of phagocytic cells isolated from the subject, an analyte isolated from a population of phagocytic cells having a DNA content more than 2n (>2n phagocytic cells) isolated from the subject, an analyte isolated from a population of circulating vesicles isolated from the subject, and an analyte isolated from a population of circulating diseased cells isolated from the subject; and

b) identifying a difference between the first profile and a second profile of at least one of the one or more markers from a repository of said markers of said disease or condition , wherein the difference is indicative of the risk of developing said disease or condition in the subject.

24. A method for prognosing or aiding in the prognosis of a disease or condition in a subject comprising:

a) determining a first profile of one or more markers of the disease or condition from a sample comprising two or more different components selected from the group consisting of: an analyte isolated from a cell-free bodily fluid isolated from the subject, an analyte isolated from a population of phagocytic cells isolated from the subject, an analyte isolated from a population of phagocytic cells having a DNA content more than 2n (>2n phagocytic cells) isolated from the subject, an analyte isolated from a population of circulating vesicles isolated from the subject, and a population of circulating diseased cells isolated from the subject; and

b) identifying a difference between the first profile and a second profile of at least one of the one or more markers from a repository of said markers of said disease or condition , wherein the difference is indicative of the prognosis of said disease or condition in the subject.

25. The method of any one of claims 1-3 and 10-12, wherein at least one of the one or more markers is up-regulated or activated in the sample compared to the control.

26. The method of any one of claims 1-3 and 10-12, wherein at least one of the one or more markers is down-regulated or inhibited in the sample compared to the control.

27. The method of any one of claims 4-6 and 13-15, wherein at least one of the one or more markers is up-regulated or activated in the first sample compared to the first control or in the second sample compared to the second control.

28. The method of any one of claims 4-6 and 13-15, wherein at least one of the one or more markers is down-regulated or inhibited in the first sample compared to the first control or in the second sample compared to the second control.

29. The method of any one of claims 7-9 and 16-18, wherein at least one of the one or more markers is up-regulated or activated in the first sample compared to the second sample.

30. The method of any one of claims 7-9 and 16-18, wherein at least one of the one or more markers is down-regulated or inhibited in the first sample compared to the second sample.

31. The method of any one of claims 19-24, wherein at least one of the one or more markers is up-regulated or activated in the sample compared to the repository.

32. The method of any one of claims 19-24, wherein at least one of the one or more markers is down-regulated or inhibited in the sample compared to the repository.

33. The method of any one of claims 1-32, wherein the first profile or the second profile comprises the absence of at least one of the one or more markers of said disease or condition.

34. The method of any one of claims 4-6 and 13-15, wherein the third profile or the fourth profile comprises the absence of at least one of the one or more markers of said disease or condition.

35. The method of any one of claims 1-9 and 19-21, wherein, when said method comprises circulating diseased cells, control cells not affected by the disease or condition, phagocytic cells, >2n phagocytic cells, =2n phagocytic cells, or non-phagocytic cells, the method further comprises lysing the circulating diseased cells, control cells not affected by the disease or condition, phagocytic cells, >2n phagocytic cells, =2n phagocytic cells, or non-phagocytic cells.

36. The method of any one of claims 1-9, 19-21, and 35, wherein, when said method comprises circulating diseased cells, control cells not affected by the disease or condition, phagocytic cells, >2n phagocytic cells, =2n phagocytic cells, or non-

phagocytic cells, the method further comprises extracting at least some of the cellular contents from the circulating diseased cells, control cells not affected by the disease or condition, phagocytic cells, >2n phagocytic cells, =2n phagocytic cells, or non-phagocytic cells.

37. The method of any one of claims 1-9, 19-21, 35, and 36, wherein, when said method comprises the cell-free bodily fluid, the method further comprises extracting the one or more markers from the cell-free bodily fluid.

38. The method of any one of claims 1-9,19-21, wherein, when said method comprises the cell-free bodily fluid, the cell-free bodily fluid comprises a transrenal nucleic acid.

39. The method of any one of claims 1-38, wherein at least one of the one or more markers of said disease or condition is present in the circulating diseased cells, cell-free bodily fluid sample, phagocytic cells, or >2n phagocytic cells.

40. The method of any one of claims 1-39, wherein at least one of the one or more markers of said disease or condition is not present in the circulating diseased cells, cell-free bodily fluid sample, phagocytic cells, or >2n phagocytic cells.

41. The method of any one of claims 1-40, wherein one or more of the circulating diseased cells, control cells not affected by the disease or condition, phagocytic cells, >2n phagocytic cells, =2n phagocytic cells, or non-phagocytic cells are enucleated.

42. The method of claim 41, wherein the cells are enucleated using physical removal, chemical treatments, photoablation, or ultraviolet irradiation.

43. The method of claim 42, wherein the physical removal uses a microneedle, optical tweezer or aspiration.

44. The method of any one of claims 19-24, wherein the repository is obtained by data mining.

45. The method of any one of claims 1-44, wherein the phagocytic cells, >2n phagocytic cells, or =2n phagocytic cells are neutrophils, macrophages, monocytes, dendritic cells, foam cells, mast cells, eosinophils, keratinocytes, or mixtures thereof.

46. The method of any one of claims 1-45, wherein the non-phagocytic cells are T cells, B cells, null cells, basophils, or mixtures thereof.

47. The method of any one of claims 1-46, wherein the circulating diseased cells are blood cells, tumor cells, lymphoma cells, fetal cells, apoptotic cells, epithelia cells,

endothelial cells, stem cells, progenitor cells, mesenchymal cells, osteoblast cells, osteocytes, hematopoietic stem cells, foam cells, adipose cells, transcervical cells, circulating cardiocytes, circulating fibrocytes, circulating myocytes, circulating cells from kidney, circulating cells from gastrointestinal tract, circulating cells from lung, circulating cells from reproductive organs, circulating cells from central nervous system, circulating hepatic cells, circulating cells from spleen, circulating cells from thymus, circulating cells from thyroid , circulating cells from an endocrine gland, circulating cells from parathyroid, circulating cells from pituitary, circulating cells from adrenal gland, circulating cells from islets of Langerhans, circulating cells from pancreas, circulating cells from hypothalamus, circulating cells from prostate tissues, circulating cells from breast tissues, circulating cells from circulating retinal cells, circulating ophthalmic cells, circulating auditory cells, circulating epidermal cells, circulating cells from the urinary tract, or mixtures thereof.

48. The method of any one of claims 1-47, wherein the control cells are normal cells.

49. The method of any one of claims 1-48, wherein the control cells are circulating cells.

50. The method of any one of claims 1-49, wherein the circulating vesicles are selected from the group consisting of circulating microvesicles, apoptotic bodies, micro-particles, membrane-bound vesicles, multivesicular bodies, nanovesicles, microparticles, and ARRDC-1 mediated microvesicles (ARMM).

51. The method of claim 50, wherein the circulating microvesicles are exosomes or urinary exosomes.

52. The method of any one of claims 1-51, wherein the cell-free bodily fluid is separated from a bodily fluid.

53. The method of claim 52, wherein the bodily fluid is blood, urine, stool, saliva, lymph fluid, cerebrospinal fluid, synovial fluid, cystic fluid, ascites, pleural effusion, fluid obtained from a pregnant woman in the first trimester, fluid obtained from a pregnant woman in the second trimester, fluid obtained from a pregnant woman in the third trimester, maternal blood, amniotic fluid, chorionic villus sample, fluid from a preimplantation embryo, maternal urine, maternal saliva, placental sample, fetal

blood, lavage and cervical vaginal fluid, interstitial fluid, buccal swab sample, sputum, bronchial lavage, Pap smear sample, or ocular fluid.

54. The method of any one of claims 52 and 53, wherein the cell-free bodily fluid is separated by filtration, centrifugation, flow cytometry, fluorescence activated cell sorting, gradient-based centrifugation, elution, microfluidics, magnetic separation technique, fluorescent-magnetic separation technique, nanostructure, quantum dots, high throughput microscope-based platforms, or a combination thereof.

55. The method of any one of claims 52-54, wherein the cell-free bodily fluid is separated by using a substance present in the sample.

56. The method of claim 55, wherein the substance is a product of a marker of said disease or condition.

57. The method of any one of claims 1-56, wherein the cell-free bodily fluid is plasma or serum.

58. The method of any one of claims 1-57, wherein the phagocytic cells, >2n phagocytic cells, =2n phagocytic cells, or non-phagocytic cells are isolated from a bodily fluid, tissues, or cells of the subject.

59. The method of claim 58, wherein the bodily fluid sample is blood, urine, stool, saliva, lymph fluid, cerebrospinal fluid, synovial fluid, cystic fluid, ascites, pleural effusion, fluid obtained from a pregnant woman in the first trimester, fluid obtained from a pregnant woman in the second trimester, fluid obtained from a pregnant woman in the third trimester, maternal blood, amniotic fluid, chorionic villus sample, fluid from a preimplantation embryo, maternal urine, maternal saliva, placental sample, fetal blood, lavage and cervical vaginal fluid, interstitial fluid, buccal swab sample, sputum, bronchial lavage, Pap smear sample, or ocular fluid.

60. The method of claim 58, wherein the cells are white blood cells.

61. The method of any one of claims 58-60, wherein the phagocytic cells, >2n phagocytic cells, =2n phagocytic cells, or non-phagocytic cells are isolated using antibodies.

62. The method of any one of claims 58-60, wherein the phagocytic cells, >2n phagocytic cells, =2n phagocytic cells, or non-phagocytic cells are isolated by flow cytometry, fluorescence activated cell sorting, filtration, gradient-based centrifugation, elution, microfluidics, magnetic separation technique, fluorescent-

magnetic separation technique, nanostructure, quantum dots, high throughput microscope-based platforms, or a combination thereof.

63. The method of any one of claims 1-62, wherein the one or more markers are nucleic acids, proteins, lipids, carbohydrates, metabolites, or combinations thereof.

64. The method of claim 63, wherein the nucleic acids are nucleotides, oligonucleotides, DNAs, RNAs, or DNA-RNA hybrids.

65. The method of claim 64, wherein the DNAs are double-stranded DNAs, single-stranded DNAs, multi-stranded DNAs, complementary DNAs, genomic DNAs, or non-coding DNAs.

66. The method of claim 64, wherein the RNAs are messenger RNAs (mRNAs), microRNAs (miRNAs), small nucleolar RNAs (snoRNAs), ribosomal RNAs (rRNAs), transfer RNAs (tRNAs), small interfering RNAs (siRNAs), heterogeneous nuclear RNAs (hnRNAs), or small hairpin RNAs (shRNAs).

67. The method of claim 63, wherein the proteins are amino acids, peptides, enzymes, antigens, antibodies, cytokines, lipoproteins, glycoproteins, or hormones.

68. The method of claim 63, wherein the lipids are fatty acids, neutral fats, phosphatides, cholesterol, cholesterol esters, triglycerides, glycolipids, glycerolipids, glycerophospholipids, sphingolipids, sterol lipids, prenol lipids, saccharolipids, polyketides, choline glycerophospholipid, ethanolamine glycerophospholipid, phosphatidylinositol, phosphatidylglycerol, phosphatidylserine, lyso-choline glycerophospholipid, lyso-ethanolamine glycerophospholipid, phosphatidic acid, lyso-phosphatidic acid, sphingomyelin, galactosylceramide, glucosylceramide, free fatty acids, prostaglandins, triacylglycerol, diacylglycerol, monoacylglycerol, acyl-CoA, acylcarnitine, oxysterol, ceramide, cardiolipin, sphingoid base-1-phosphate, shingosine, lyso-sphingomyelin,, gangliosides, plasmalogen, sulfatide, low density lipoproteins (LDLs), very low density lipoproteins (VLDLs), high density lipoproteins (HDLs), sphingoid base-1-phosphates, or derivatives thereof.

69. The method of claim 63, wherein the carbohydrates are monosaccharides, disaccharides, polysaccharides, oligosaccharides, or derivatives thereof.

70. The method of claim 63, wherein the metabolites are primary metabolites, secondary metabolites, organic metabolites, inorganic metabolites, prostaglandins,

hydroxyeicosatetraenoic acids, hydroxyoctadecadienoic acids, steroids, bile acids, vitamins, or derivatives thereof.

71. The method of any one of claims 1-3, 10-12, and 19-24, wherein the profile is a nucleic acid profile, a protein profile, a lipid profile, a carbohydrate profile, a metabolite profile, or a combination thereof.

72. The method of claim 65, wherein the profile is determined by a qualitative assay, a quantitative assay, or a combination thereof.

73. The method of any one of claims 4-9 and 13-18, wherein the first profile or the second profile is a nucleic acid profile, a protein profile, a lipid profile, a carbohydrate profile, a metabolite profile, or a combination thereof.

74. The method of claim 65, wherein the first profile or the second profile is determined by a qualitative assay, a quantitative assay, or a combination thereof.

75. The method of any one of claims 4-6 and 74, wherein the third profile or the fourth profile is a nucleic acid profile, a protein profile, a lipid profile, a carbohydrate profile, a metabolite profile, or a combination thereof.

76. The method of claim 75, wherein the third profile or the fourth profile is determined by a qualitative assay, a quantitative assay, or a combination thereof.

77. The method of any one of claims 72, 66, and 76, wherein the quantitative assay uses sequencing, targeted sequencing, single molecule real-time sequencing, electron microscopy-based sequencing, transistor-mediated sequencing, direct sequencing, random shotgun sequencing, Sanger dideoxy termination sequencing, exon sequencing, whole-genome sequencing, sequencing by hybridization, pyrosequencing, capillary electrophoresis, gel electrophoresis, duplex sequencing, cycle sequencing, single-base extension sequencing, solid-phase sequencing, high-throughput sequencing, massively parallel signature sequencing, emulsion PCR, multiplex PCR, co-amplification at lower denaturation temperature-PCR (COLD-PCR), sequencing by reversible dye terminator, paired-end sequencing, near-term sequencing, exonuclease sequencing, sequencing by ligation, short-read sequencing, single-molecule sequencing, sequencing-by-synthesis, real-time sequencing, reverse-terminator sequencing, nanopore sequencing, 454 sequencing, Solexa Genome Analyzer sequencing, SOLiD® sequencing, MS-PET sequencing, mass spectrometry, matrix assisted laser desorption/ionization-time of flight (MALDI-TOF) mass

spectrometry, electrospray ionization (ESI) mass spectrometry, surface-enhanced laser deorption/ionization-time of flight (SELDI-TOF) mass spectrometry, quadrupole-time of flight (Q-TOF) mass spectrometry, atmospheric pressure photoionization mass spectrometry (APPI-MS), Fourier transform mass spectrometry (FTMS), matrix-assisted laser desorption/ionization-Fourier transform-ion cyclotron resonance (MALDI-FT-ICR) mass spectrometry, secondary ion mass spectrometry (SIMS), polymerase chain reaction (PCR) analysis, co-amplification at lower denaturation temperature-PCR (COLD-PCR), multiplex PCR, quantitative PCR, real-time PCR, fluorescence assay, colorimetric assay, chemiluminescent assay, or a combination thereof.

78. The method of claim 77, wherein the nucleic acid profile is a genotypic profile, a single nucleotide polymorphism profile, a gene mutation profile, a gene copy number profile, a DNA methylation profile, a DNA acetylation profile, a chromosome dosage profile, a gene expression profile, or a combination thereof.

79. The method of claim 77, wherein the nucleic acid profile is determined by polymerase chain reaction (PCR) analysis, sequencing analysis, electrophoretic analysis, restriction fragment length polymorphism (RFLP) analysis, Northern blot analysis, quantitative PCR, reverse-transcriptase-PCR analysis (RT-PCR), allele-specific oligonucleotide hybridization analysis, comparative genomic hybridization, heteroduplex mobility assay (HMA), single strand conformational polymorphism (SSCP), denaturing gradient gel electrophisis (DGGE), RNAase mismatch analysis, mass spectrometry, tandem mass spectrometry, matrix assisted laser desorption/ionization-time of flight (MALDI-TOF) mass spectrometry, electrospray ionization (ESI) mass spectrometry, surface-enhanced laser deorption/ionization-time of flight (SELDI-TOF) mass spectrometry, quadrupole-time of flight (Q-TOF) mass spectrometry, atmospheric pressure photoionization mass spectrometry (APPI-MS), Fourier transform mass spectrometry (FTMS), matrix-assisted laser desorption/ionization-Fourier transform-ion cyclotron resonance (MALDI-FT-ICR) mass spectrometry, secondary ion mass spectrometry (SIMS), surface plasmon resonance, Southern blot analysis, in situ hybridization, fluorescence in situ hybridization (FISH), chromogenic in situ hybridization (CISH), immunohistochemistry (IHC), microarray, comparative genomic hybridization,

karyotyping, multiplex ligation-dependent probe amplification (MLPA), Quantitative Multiplex PCR of Short Fluorescent Fragments (QMPSF), microscopy, methylation specific PCR (MSP) assay, HpaII tiny fragment Enrichment by Ligation-mediated PCR (HELP) assay, radioactive acetate labeling assays, colorimetric DNA acetylation assay, chromatin immunoprecipitation combined with microarray (ChIP-on-chip) assay, restriction landmark genomic scanning, Methylated DNA immunoprecipitation (MeDIP), molecular break light assay for DNA adenine methyltransferase activity, chromatographic separation, methylation-sensitive restriction enzyme analysis, bisulfite-driven conversion of non-methylated cytosine to uracil, co-amplification at lower denaturation temperature-PCR (COLD-PCR), multiplex PCR, methyl-binding PCR analysis, or a combination thereof.

80. The method of claim 77, wherein the nucleic acid profile is determined by a sequencing technique selected from the group consisting of targeted sequencing, single molecule real-time sequencing, exon sequencing, electron microscopy-based sequencing, transistor-mediated sequencing, direct sequencing, random shotgun sequencing, Sanger dideoxy termination sequencing, whole-genome sequencing, sequencing by hybridization, pyrosequencing, capillary electrophoresis, gel electrophoresis, duplex sequencing, cycle sequencing, single-base extension sequencing, solid-phase sequencing, high-throughput sequencing, massively parallel signature sequencing, emulsion PCR, co-amplification at lower denaturation temperature-PCR (COLD-PCR), multiplex PCR, sequencing by reversible dye terminator, paired-end sequencing, near-term sequencing, exonuclease sequencing, sequencing by ligation, short-read sequencing, single-molecule sequencing, sequencing-by-synthesis, real-time sequencing, reverse-terminator sequencing, nanopore sequencing, 454 sequencing, Solexa Genome Analyzer sequencing, SOLiD® sequencing, MS-PET sequencing, mass spectrometry, and a combination thereof.

81. The method of claim 77, wherein the protein profile is a protein expression profile, a protein activation profile, or a combination thereof.

82. The method of claim 77, wherein the protein profile is determined by an immunohistochemistry assay, an enzyme-linked immunosorbent assay (ELISA), in situ hybridization, chromatography, liquid chromatography, size exclusion

chromatography, high performance liquid chromatography (HPLC), gas chromatography, mass spectrometry, tandem mass spectrometry, matrix assisted laser desorption/ionization-time of flight (MALDI-TOF) mass spectrometry, electrospray ionization (ESI) mass spectrometry, surface-enhanced laser deorption/ionization-time of flight (SELDI-TOF) mass spectrometry, quadrupole-time of flight (Q-TOF) mass spectrometry, atmospheric pressure photoionization mass spectrometry (APPI-MS), Fourier transform mass spectrometry (FTMS), matrix-assisted laser desorption/ionization-Fourier transform-ion cyclotron resonance (MALDI-FT-ICR) mass spectrometry, secondary ion mass spectrometry (SIMS), radioimmunoassays, microscopy, microfluidic chip-based assays, surface plasmon resonance, sequencing, Western blotting assay, or a combination thereof.

83. The method of claim 77, wherein the protein activation profile comprises determining a phosphorylation state, an ubiquitination state, a myristoylation state, a conformational state, or a combination thereof of the one or more markers.

84. The method of claim 77, wherein the lipid profile is determined by chromatography, liquid chromatography, size exclusion chromatography, high performance liquid chromatography (HPLC), gas chromatography, mass spectrometry, tandem mass spectrometry, matrix assisted laser desorption/ionization-time of flight (MALDI-TOF) mass spectrometry, electrospray ionization (ESI) mass spectrometry, surface-enhanced laser deorption/ionization-time of flight (SELDI-TOF) mass spectrometry, quadrupole-time of flight (Q-TOF) mass spectrometry, atmospheric pressure photoionization mass spectrometry (APPI-MS), Fourier transform mass spectrometry (FTMS), matrix-assisted laser desorption/ionization-Fourier transform-ion cyclotron resonance (MALDI-FT-ICR) mass spectrometry, secondary ion mass spectrometry (SIMS), radioimmunoassays, microfluidic chip-based assay, detection of fluorescence, detection of chemiluminescence, or a combination thereof.

85. The method of claim 77, wherein the carbohydrate profile is determined by chromatography, liquid chromatography, size exclusion chromatography, high performance anion exchange chromatography with pulsed amperometric detection (HPAEC-PAD), liquid chromatography, gas chromatography, fluorescent assay, mass spectrometry, tandem mass spectrometry, matrix assisted laser desorption/ionization-

time of flight (MALDI-TOF) mass spectrometry, electrospray ionization (ESI) mass spectrometry, surface-enhanced laser deorption/ionization-time of flight (SELDI-TOF) mass spectrometry, quadrupole-time of flight (Q-TOF) mass spectrometry, atmospheric pressure photoionization mass spectrometry (APPI-MS), Fourier transform mass spectrometry (FTMS), matrix-assisted laser desorption/ionization-Fourier transform-ion cyclotron resonance (MALDI-FT-ICR) mass spectrometry, secondary ion mass spectrometry (SIMS), radioimmunoassay, microfluidic chip-based assay, detection of fluorescence, detection of chemiluminescence, or a combination thereof.

86. The method of any one of claims 1-85, wherein the subject has at least two diseases or conditions.

87. The method of claim 86, wherein the subject has at least one prenatal or pregnancy-related disease or condition

88. The method of any one of claims 1-87, wherein the subject is a mammal.

89. The method of claim 89, wherein the mammal is a human.

90. The method of any one of claims 1-89, wherein the disease or condition is a cardiovascular disease or condition, a kidney-associated disease or condition, a prenatal or pregnancy-related disease or condition, a neurological or neuropsychiatric disease or condition, an autoimmune or immune-related disease or condition, a cancer, an infectious disease or condition, a mitochondrial disorder, a respiratory-gastrointestinal tract disease or condition, a reproductive disease or condition, an ophthalmic disease or condition, a musculo-skeletal disease or condition, or a dermal disease or condition.

91. The method of any one of claims 1-90, wherein the difference is greater than a 1-fold difference.

92. The method of claim 91, wherein the difference is at least 1.05-fold, 1.1-fold, 1.2-fold, 1.3-fold, 1.4-fold, 1.5-fold, 2-fold, 2.5-fold, 3-fold, 4-fold, 5-fold, 6-fold, 7-fold, 8-fold, 9-fold, or 10-fold difference.

93. The method of any one of claims 4-6 and 13-15, wherein the second profile and the fourth profile are the same.

94. A method for identifying one or more markers for a disease or condition comprising:

a) determining a first profile of analytes from a sample comprising a cell-free bodily fluid from a subject having said disease or condition and a population of phagocytic cells, or a population of >2n phagocytic cells, from a subject having said disease or condition;

determining a second profile of analytes from a population of =2n phagocytic cells, or a population of non-phagocytic cells, from the subject having said disease or condition;

identifying a set of differences between the first and second profiles, wherein the first set of differences is specific to the first profile relative to the second profile;

b) determining a third profile of analytes from a sample comprising a cell-free bodily fluid from a subject having said disease or condition and a population of phagocytic cells, or a population of >2n phagocytic cells, from a control subject not having said disease or condition;

determining a fourth profile of analytes from a population of =2n phagocytic cells, or a population of non-phagocytic cells, from the control subject not having said disease or condition;

identifying a set of differences between the third and fourth profiles, wherein the second set of differences is specific to the third profile relative to the fourth profile; and

c) identifying one or more analytes specific to the set of differences identified in a) relative to the set of differences identified in b), the identified analytes in c) being markers of said disease or condition.

95. A method for identifying one or more markers for a disease or condition comprising:

a) determining a first profile of analytes from a sample comprising a cell-free bodily fluid from a subject having said disease or condition and a population of phagocytic cells, or a population of >2n phagocytic cells, from a subject having said disease or condition;

b) comparing the first profile to a second profile derived from a repository of analytes from a control subject not having said disease or condition;

c) identifying a set of differences between the first and second profiles, wherein the set of differences is specific to the first profile relative to the second profile; and

d) identifying one or more analytes specific to the set of differences, the identified analytes being markers of said disease or condition.

96. The method of claim 95, further comprising:

a) obtaining a fifth profile of analytes from cells or tissues affected by said disease or condition in the subject having said disease or condition;

obtaining a sixth profile of analytes from cells or tissues not affected by said disease or condition in the subject having said disease or condition;

identifying a set of differences between the fifth and sixth profiles, wherein the set of differences is specific to the fifth profile relative to the sixth profile; and

b) identifying at least one of the one or more markers of c) present in the set of differences identified in d).

97. The method of any one of claims 90-96, further comprising extracting the one or more markers from the cell-free bodily fluid.

98. The method of any one of claims 90-97, further comprising lysing the phagocytic cells, the >2n phagocytic cells, the =2n phagocytic cells, or the non-phagocytic cells before a).

99. The method of any one of claims 90-98, further comprising extracting at least some of the cellular contents from the phagocytic cells, the >2n phagocytic cells, the =2n phagocytic cells, or the non-phagocytic cells before a).

100. The method of any one of claims 90-99, wherein the phagocytic cells or the >2n phagocytic cells comprise viable diseased cells, dead diseased cells, apoptotic diseased cells, circulating tumor cells, infectious agents, fetal cells, trophoblasts, or fragments thereof.

101. The method of any one of claims 90-100, wherein at least one of the one or more markers of said disease or condition is present in the cell-free bodily fluid sample, the phagocytic cells, or the >2n phagocytic cells.

102. The method of any one of claims 90-101, wherein at least one of the one or more markers of said disease or condition is not present in the cell-free bodily fluid sample, the phagocytic cells, or the >2n phagocytic cells.

103. The method of any one of claims 90-102, further comprising comparing the identified difference of c) to a repository of one or more known markers of said disease or condition.

104. The method of claim 97, wherein the repository is obtained by data mining.

105. The method of any one of claims 90-104, wherein the phagocytic cells, >2n phagocytic cells, or =2n phagocytic cells are neutrophils, macrophages, monocytes, dendritic cells, foam cells, mast cells, eosinophils, keratinocytes, or mixtures thereof.

106. The method of any one of the claims 90-105, wherein the non-phagocytic cells are T cells, B cells, null cells, basophils, or mixtures thereof

107. The method of any one of claims 90-106, wherein the cell-free bodily fluid is separated from a bodily fluid.

108. The method of claim 107, wherein the bodily fluid sample is blood, urine, stool, saliva, lymph fluid, cerebrospinal fluid, synovial fluid, cystic fluid, ascites, pleural effusion, fluid obtained from a pregnant woman in the first trimester, fluid obtained from a pregnant woman in the second trimester, fluid obtained from a pregnant woman in the third trimester, maternal blood, amniotic fluid, chorionic villus sample, fluid from a preimplantation embryo, maternal urine, maternal saliva, placental sample, fetal blood, lavage and cervical vaginal fluid, interstitial fluid, or ocular fluid.

109. The method of any one of claims 107 and 108, wherein the cell-free bodily fluid sample is separated by filtration, centrifugation, flow cytometry, fluorescence activated cell sorting, gradient-based centrifugation, elution, microfluidics, magnetic separation technique, fluorescent-magnetic separation technique, nanostructure, quantum dots, high throughput microscope-based platforms, or a combination thereof.

110. The method of claim 109, wherein the cell-free bodily fluid sample is separated by using a substance present in the sample.

111. The method of any one of claims 90-110, wherein the one or more markers are nucleic acids, proteins, lipids, carbohydrates, metabolites, or combinations thereof.

112. The method of claim 111, wherein the nucleic acids are nucleotides, oligonucleotides, DNAs, RNAs, or DNA-RNA hybrids.

113. The method of claim 112, wherein the DNAs are double-stranded DNAs, single-stranded DNAs, multi-stranded DNAs, complementary DNAs, genomic DNAs, or non-coding DNAs.

114. The method of claim 112, wherein the RNAs are messenger RNAs (mRNAs), microRNAs (miRNAs), small nucleolar RNAs (snoRNAs), ribosomal RNAs (rRNAs), transfer RNAs (tRNAs), small interfering RNAs (siRNAs), heterogeneous nuclear RNAs (hnRNAs), or small hairpin RNAs (shRNAs).

115. The method of claim 111, wherein the proteins are amino acids, peptides, enzymes, antigens, antibodies, cytokines, lipoproteins, glycoproteins, or hormones.

116. The method of claim 111, wherein the lipids are fatty acids, neutral fats, phosphatides, cholesterol, cholesterol esters, triglycerides, glycolipids, glycerolipids, glycerophospholipids, sphingolipids, sterol lipids, prenol lipids, saccharolipids, polyketides, choline glycerophospholipid, ethanolamine glycerophospholipid, phosphatidylinositol, phosphatidylglycerol, phosphatidylserine, lyso-choline glycerophospholipid, lyso-ethanolamine glycerophospholipid, phosphatidic acid, lyso-phosphatidic acid, sphingomyelin, galactosylceramide, glucosylceramide, free fatty acids, prostaglandins, triacylglycerol, diacylglycerol, monoacylglycerol, acyl-CoA, acylcarnitine, oxysterol, ceramide, cardiolipin, sphingoid base-1-phosphate, shingosine, lyso-sphingomyelin,, gangliosides, plasmalogen, sulfatide, low density lipoproteins (LDLs), very low density lipoproteins (VLDLs), high density lipoproteins (HDLs), sphingoid base-1-phosphates, or derivatives thereof.

117. The method of claim 111, wherein the carbohydrates are monosaccharides, disaccharides, polysaccharides, oligosaccharides, or derivatives thereof.

118. The method of claim 111, wherein the metabolites are primary metabolites, secondary metabolites, organic metabolites, inorganic metabolites, prostaglandins, hydroxyeicosatetraenoic acids, hydroxyoctadecadienoic acids, steroids, bile acids, vitamins, or derivatives thereof.

119. The method of any one of claims 90-118, wherein the profile is a nucleic acid profile, a protein profile, a lipid profile, a carbohydrate profile, a metabolite profile, or a combination thereof.

120. The method of claim 119, wherein the profile is determined by a qualitative assay, a quantitative assay, or a combination thereof.

121. The method of claim 120, wherein the quantitative assay uses sequencing, targeted sequencing, single molecule real-time sequencing, exon sequencing, electron microscopy-based sequencing, transistor-mediated sequencing, direct sequencing, random shotgun sequencing, Sanger dideoxy termination sequencing, whole-genome sequencing, sequencing by hybridization, pyrosequencing, capillary electrophoresis, gel electrophoresis, duplex sequencing, cycle sequencing, single-base extension sequencing, solid-phase sequencing, high-throughput sequencing, massively parallel signature sequencing, emulsion PCR, co-amplification at lower denaturation temperature-PCR (COLD-PCR), multiplex PCR, sequencing by reversible dye terminator, paired-end sequencing, near-term sequencing, exonuclease sequencing, sequencing by ligation, short-read sequencing, single-molecule sequencing, sequencing-by-synthesis, real-time sequencing, reverse-terminator sequencing, nanopore sequencing, 454 sequencing, Solexa Genome Analyzer sequencing, SOLiD® sequencing, MS-PET sequencing, mass spectrometry, matrix assisted laser desorption/ionization-time of flight (MALDI-TOF) mass spectrometry, electrospray ionization (ESI) mass spectrometry, surface-enhanced laser deorption/ionization-time of flight (SELDI-TOF) mass spectrometry, quadrupole-time of flight (Q-TOF) mass spectrometry, atmospheric pressure photoionization mass spectrometry (APPI-MS), Fourier transform mass spectrometry (FTMS), matrix-assisted laser desorption/ionization-Fourier transform-ion cyclotron resonance (MALDI-FT-ICR) mass spectrometry, secondary ion mass spectrometry (SIMS), polymerase chain reaction (PCR) analysis, quantitative PCR, real-time PCR, fluorescence assay, colorimetric assay, chemiluminescent assay, or a combination thereof.

122. The method of claim 119, wherein the nucleic acid profile is a genotypic profile, a single nucleotide polymorphism profile, a gene mutation profile, a gene copy number profile, a DNA methylation profile, a DNA acetylation profile, a chromosome dosage profile, a gene expression profile, or a combination thereof.

123. The method of claim 119, wherein the nucleic acid profile is determined by polymerase chain reaction (PCR) analysis, sequencing analysis, electrophoretic analysis, restriction fragment length polymorphism (RFLP) analysis, Northern blot

analysis, quantitative PCR, reverse-transcriptase-PCR analysis (RT-PCR), allele-specific oligonucleotide hybridization analysis, comparative genomic hybridization, heteroduplex mobility assay (HMA), single strand conformational polymorphism (SSCP), denaturing gradient gel electrophisis (DGGE), RNAase mismatch analysis, mass spectrometry, tandem mass spectrometry, matrix assisted laser desorption/ionization-time of flight (MALDI-TOF) mass spectrometry, electrospray ionization (ESI) mass spectrometry, surface-enhanced laser deorption/ionization-time of flight (SELDI-TOF) mass spectrometry, quadrupole-time of flight (Q-TOF) mass spectrometry, atmospheric pressure photoionization mass spectrometry (APPI-MS), Fourier transform mass spectrometry (FTMS), matrix-assisted laser desorption/ionization-Fourier transform-ion cyclotron resonance (MALDI-FT-ICR) mass spectrometry, secondary ion mass spectrometry (SIMS), surface plasmon resonance, Southern blot analysis, in situ hybridization, fluorescence in situ hybridization (FISH), chromogenic in situ hybridization (CISH), immunohistochemistry (IHC), microarray, comparative genomic hybridization, karyotyping, multiplex ligation-dependent probe amplification (MLPA), Quantitative Multiplex PCR of Short Fluorescent Fragments (QMPSF), microscopy, methylation specific PCR (MSP) assay, HpaII tiny fragment Enrichment by Ligation-mediated PCR (HELP) assay, radioactive acetate labeling assays, colorimetric DNA acetylation assay, chromatin immunoprecipitation combined with microarray (ChIP-on-chip) assay, restriction landmark genomic scanning, Methylated DNA immunoprecipitation (MeDIP), molecular break light assay for DNA adenine methyltransferase activity, chromatographic separation, methylation-sensitive restriction enzyme analysis, bisulfite-driven conversion of non-methylated cytosine to uracil, co-amplification at lower denaturation temperature-PCR (COLD-PCR), multiplex PCR, methyl-binding PCR analysis, or a combination thereof.

124. The method of claim 119, wherein the nucleic acid profile is determined by a sequencing technique selected from the group consisting of targeted sequencing, single molecule real-time sequencing, exon sequencing, electron microscopy-based sequencing, transistor-mediated sequencing, direct sequencing, random shotgun sequencing, Sanger dideoxy termination sequencing, whole-genome sequencing, sequencing by hybridization, pyrosequencing, capillary electrophoresis, gel

electrophoresis, duplex sequencing, cycle sequencing, single-base extension sequencing, solid-phase sequencing, high-throughput sequencing, massively parallel signature sequencing, emulsion PCR, co-amplification at lower denaturation temperature-PCR (COLD-PCR), multiplex PCR, sequencing by reversible dye terminator, paired-end sequencing, near-term sequencing, exonuclease sequencing, sequencing by ligation, short-read sequencing, single-molecule sequencing, sequencing-by-synthesis, real-time sequencing, reverse-terminator sequencing, nanopore sequencing, 454 sequencing, Solexa Genome Analyzer sequencing, SOLiD® sequencing, MS-PET sequencing, mass spectrometry, and a combination thereof.

125. The method of claim 119, wherein the protein profile is a protein expression profile, a protein activation profile, or a combination thereof.

126. The method of claim 119, wherein the protein profile is determined by an immunohistochemistry assay, an enzyme-linked immunosorbent assay (ELISA), in situ hybridization, chromatography, liquid chromatography, size exclusion chromatography, high performance liquid chromatography (HPLC), gas chromatography, mass spectrometry, tandem mass spectrometry, matrix assisted laser desorption/ionization-time of flight (MALDI-TOF) mass spectrometry, electrospray ionization (ESI) mass spectrometry, surface-enhanced laser deorption/ionization-time of flight (SELDI-TOF) mass spectrometry, quadrupole-time of flight (Q-TOF) mass spectrometry, atmospheric pressure photoionization mass spectrometry (APPI-MS), Fourier transform mass spectrometry (FTMS), matrix-assisted laser desorption/ionization-Fourier transform-ion cyclotron resonance (MALDI-FT-ICR) mass spectrometry, secondary ion mass spectrometry (SIMS), radioimmunoassays, microscopy, microfluidic chip-based assays, surface plasmon resonance, sequencing, Western blotting assay, or a combination thereof.

127. The method of claim 119, wherein the protein activation profile comprises determining a phosphorylation state, an ubiquitination state, a myristoylation state, a conformational state, or a combination thereof of the one or more markers.

128. The method of claim 119, wherein the lipid profile is determined by chromatography, liquid chromatography, size exclusion chromatography, high performance liquid chromatography (HPLC), gas chromatography, mass

spectrometry, tandem mass spectrometry, matrix assisted laser desorption/ionization-time of flight (MALDI-TOF) mass spectrometry, electrospray ionization (ESI) mass spectrometry, surface-enhanced laser deorption/ionization-time of flight (SELDI-TOF) mass spectrometry, quadrupole-time of flight (Q-TOF) mass spectrometry, atmospheric pressure photoionization mass spectrometry (APPI-MS), Fourier transform mass spectrometry (FTMS), matrix-assisted laser desorption/ionization-Fourier transform-ion cyclotron resonance (MALDI-FT-ICR) mass spectrometry, secondary ion mass spectrometry (SIMS), radioimmunoassays, microfluidic chip-based assay, detection of fluorescence, detection of chemiluminescence, or a combination thereof.

129. The method of claim 119, wherein the carbohydrate profile is determined by chromatography, liquid chromatography, size exclusion chromatography, high performance anion exchange chromatography with pulsed amperometric detection (HPAEC-PAD), liquid chromatography, gas chromatography, fluorescent assay, mass spectrometry, tandem mass spectrometry, matrix assisted laser desorption/ionization-time of flight (MALDI-TOF) mass spectrometry, electrospray ionization (ESI) mass spectrometry, surface-enhanced laser deorption/ionization-time of flight (SELDI-TOF) mass spectrometry, quadrupole-time of flight (Q-TOF) mass spectrometry, atmospheric pressure photoionization mass spectrometry (APPI-MS), Fourier transform mass spectrometry (FTMS), matrix-assisted laser desorption/ionization-Fourier transform-ion cyclotron resonance (MALDI-FT-ICR) mass spectrometry, secondary ion mass spectrometry (SIMS), radioimmunoassay, microfluidic chip-based assay, detection of fluorescence, detection of chemiluminescence, or a combination thereof.

130. The method of any one of claims 90-129, wherein the subject is a mammal

131. The method of claim 130, wherein the mammal is a human.

132. The method of any one of claims 90-131, wherein the disease or condition is a cardiovascular disease or condition, a kidney-associated disease or condition, a prenatal or pregnancy-related disease or condition, a neurological or neuropsychiatric disease or condition, an autoimmune or immune-related disease or condition, a cancer, an infectious disease or condition, a mitochondrial disorder, a respiratory-gastrointestinal tract disease or condition, a reproductive disease or condition, an

ophthalmic disease or condition, a musculo-skeletal disease or condition, or a dermal disease or condition.

133. The method of any one of claims 90-132, wherein the difference is greater than a 1-fold difference.

134. The method of claim 133, wherein the difference is at least 1.05-fold, 1.1-fold, 1.2-fold, 1.3-fold, 1.4-fold, 1.5-fold, 2-fold, 2.5-fold, 3-fold, 4-fold, 5-fold, 6-fold, 7-fold, 8-fold, 9-fold, or 10-fold difference.

135. The method of any one of claims 1-93, further comprising determining at least one diagnostic parameter of said disease or condition.

136. The method of claim 135, wherein the diagnostic parameter is determined by physical inspection, visual inspection, biopsy, scanning, histology, radiology, imaging, ultrasound, use of a commercial kit, genetic testing, immunological testing, analysis of bodily fluids, or monitoring neural activity.

137. The method of any one of the claims 1-93 and 135-136, wherein the one or more markers comprise at least one gene selected from the group consisting of AKT2, BAK1, EGFR, ERBB2, ETS2, FOS, JUN, MAP2K1, MMP2, PDGFB, RB1, SERPINB2, SNCG, and SPP1.

138. The method of any one of the claims 1-93 and 135-137, wherein the one or more markers comprise at least one gene selected from the group consisting of AKT1, AKT2, BAK2, CDC25A, E2F1, EGFR, ERBB2, FOS, JUN, MAP2K1, MMP2, NFKB1, PDGFB, PIK3R1, PNN, RB1, SERPINB2, SERPINB5, SNCG, SPP1, TERT, TIMP3, and TP53.

139. The method of any one of the claims 1-93 and 135-138, wherein the one or more markers comprise at least one gene selected from the group consisting of CASP8, CASP9, COL18A1, ETS2, HTATIP2, MMP9, SRC, and TWIST1.

140. The method of any one of the claims 1-93 and 135-139, wherein the one or more markers comprise at least one gene selected from the group consisting of AKT1, APAF1, ATM, CDC25A, CDKN1A, ETS2, FOS, IL8, ITGA4, ITGA6, ITGAV, JUN, MAP2K1, NFKBIA, PLAU, PLAUR, RAF1, SERPINB2, SYK, TIMP1, TNF, TNFRSF10B, and TNFRSF1A.

141. The method of any one of the claims 1-93 and 135-140, wherein the one or more markers comprise at least one gene selected from the group consisting of ACP2,

AK2, AKT3, ARL5B, ATP2B3, BGN, BRAF, BTG2, CAMKK2, CAPG, CAPN12, CPLX2, DENND5A, DNA2, FAM104A, FNIP1, GFRA4, GLUD1, GNAQ, GP1BB, HNRPLL, HOXA2, HPS3, INPP4A, ITGAV, KLHL23, LANCL2, LYPD6, MAPKAPK3, MEF2A (includes, EG:4205), MEF2C, NVL, PCYT1A, PGLYRP4, PLOD1, PPP1CB, PRKAB2, PROS1, PTPRE, RASA4 (includes,EG:10156), RBMS2, RBPJ, STAT5B, THBS1, TRIB1, TRIM2, TSPAN6, and ZDHHC21.

142. The method of any one of the claims 1-93 and 135-141, wherein the one or more markers comprise at least one gene selected from the group consisting of B4GALT5, BOP1, CCL2, CCL3, CCL3L1, CCRL2, CD83, CLEC4G, CLIC4, CTSC, CTSO, CXCL10, FCGR3A, FPR3, HBA1, HBB, LRMP, MAP1LC3B2, MS4A4A, MSR1, MYADML, NID1, PF4, PION, RNF217, SAMD9L, SERPING1, and SPARC.

143. The method of any one of the claims 1-93 and 135-142, wherein the one or more markers comprise at least one gene selected from the group consisting of ACOT9, AMPD2, ARHGAP15, BATF2, C3AR1, C5orf41, CCL3, CCL3L1, CD63, CHST11, CHSY1, CLEC4G, CTSZ, CXorf21, CYTH4, CYTIP, DLEU2, DNAJA1, DOCK8, DTX3L, DUSP6, EPSTI1, ERF, F2RL1, FYB, GABRB2, GBP5, GLRX, GNB4, ICAM1, IFI35, IFIH1, IFNAR2, IL1R1, IRF1, ITGA5, LAP3, LAPTM5, LCP2, MAP1LC3B, MAP1LC3B2, MICAL2, MT1DP, MT1JP, MT1M, MT2A, MYADML, NEK6, NINJ2, NNMT, NT5C3L, NUB1, PDE4B, PLOD1, PML, PRKCB, PSMB9, RCN3, RGS4, RNASE6, RTP4, SAMD9L, SEL1L, SERPING1, SETX, SIGLEC10, SKIL, SLC7A7, SNORA21, SP100, SP110, SP140, SSFA2, STAT2, STK17B, STK3, TDRD7, TMCC1, TMPRSS11E2, TNFRSF1B, TPM1, TRIM21,TXNDC4, UBE2L6, UBE2W, USP18, VAV1, WARS, WIPF1, and WIPI1.

144. The method of any one of the claims 1-93 and 135-143, wherein the one or more markers comprise at least one gene selected from the group consisting of ADAR, ADM, ALAS1, ANKRD22, ARHGAP27, B3GNT5, BCL10, C12orf35, C15orf29, C2orf59, CD177, CEACAM1, CPEB2, DDX58, F2RL1, GDPD3, GNAI3, HIST2H3A, HIST2H3D, HIST2H4A, HMGCR, HSPA6, HSPC159, IL4R, IMPA2, KPNB1, KREMEN1, KRT23, LDLR, LOC100130904, LTB4R, MAEA, MARK2, MBOAT2, MPZL3, N4BP1, NBEAL2, NMI, NPEPPS, PARP14, PGM2, PPIF, PXN, RALBP1, ROD1, RPS6KA1, S100P, SERTAD2, SLC9A1, SLPI, SP110, SPINT1, ST14, TBC1D3, TNFRSF9, TRIM21, UPP1, VPS24, ZBTB34, and ZNF256.

145. The method of any one of the claims 1-93 and 135-144, wherein the one or more markers comprise at least one or more of the markers identified by the methods of any one the claims 94-134.

146. A kit comprising a plurality of marker detection agents that detect at least one or more of the markers identified by the methods of any one of the claims 94-134.

147. A method of treating or preventing a disease or condition in a subject comprising administering to said subject a composition comprising a compound identified by the method of any one of claims 6 and 15.

148. The method of any one of the claims 1-93 and 135-145, wherein the circulating diseased cells are infected by an infectious agent.

149. The method of claim 148, wherein the infectious agent is a virus, bacteria, fungus, parasite, protozoan, infectious protein or microorganism.

150. The method of any one of claims 1-93, 135-145, and 148-149, wherein, when said method comprises phagocytic cells or >2n phagocytic cells, said phagocytic cells or >2n phagocytic cells comprise a transrenal nucleic acid.

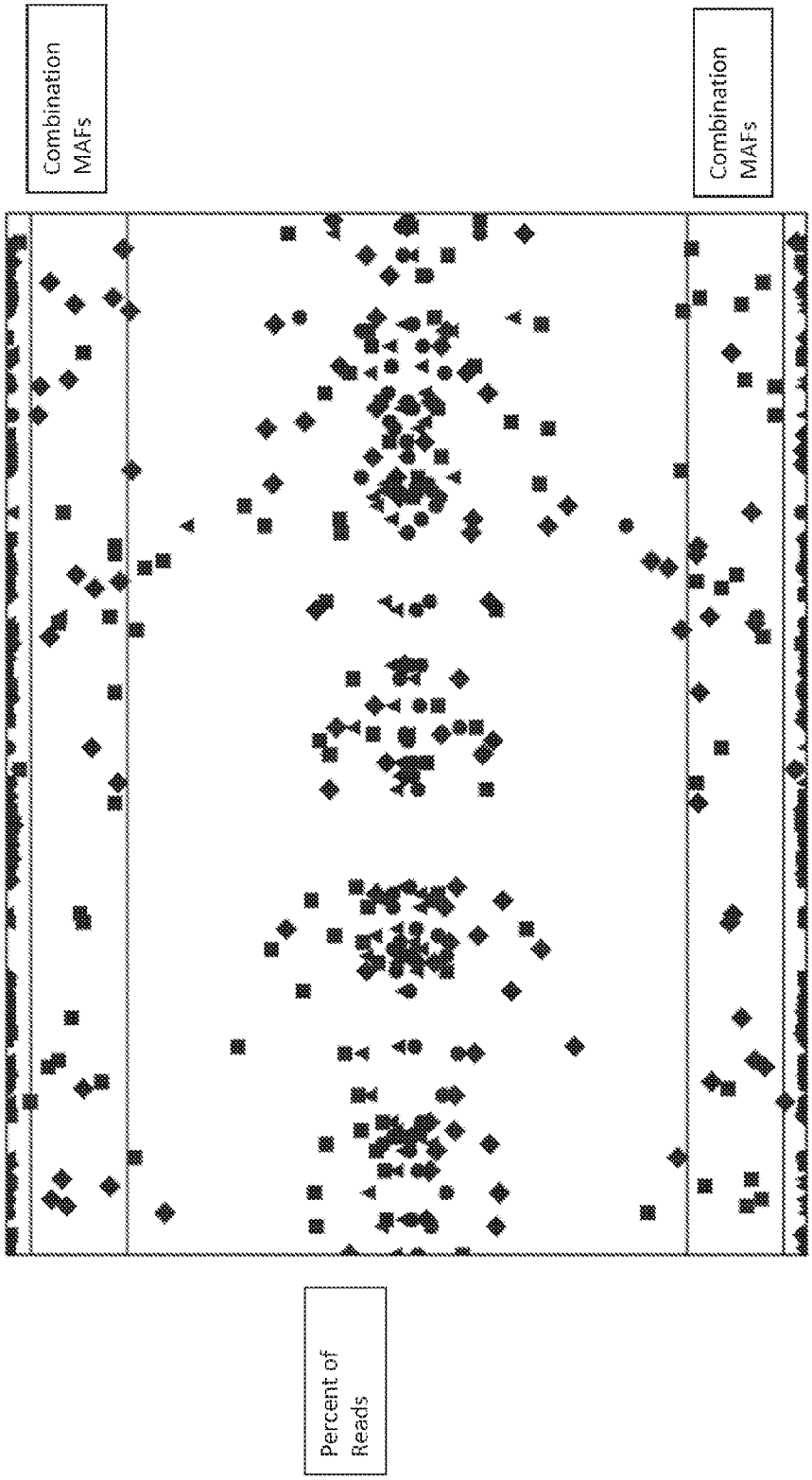


Figure 1

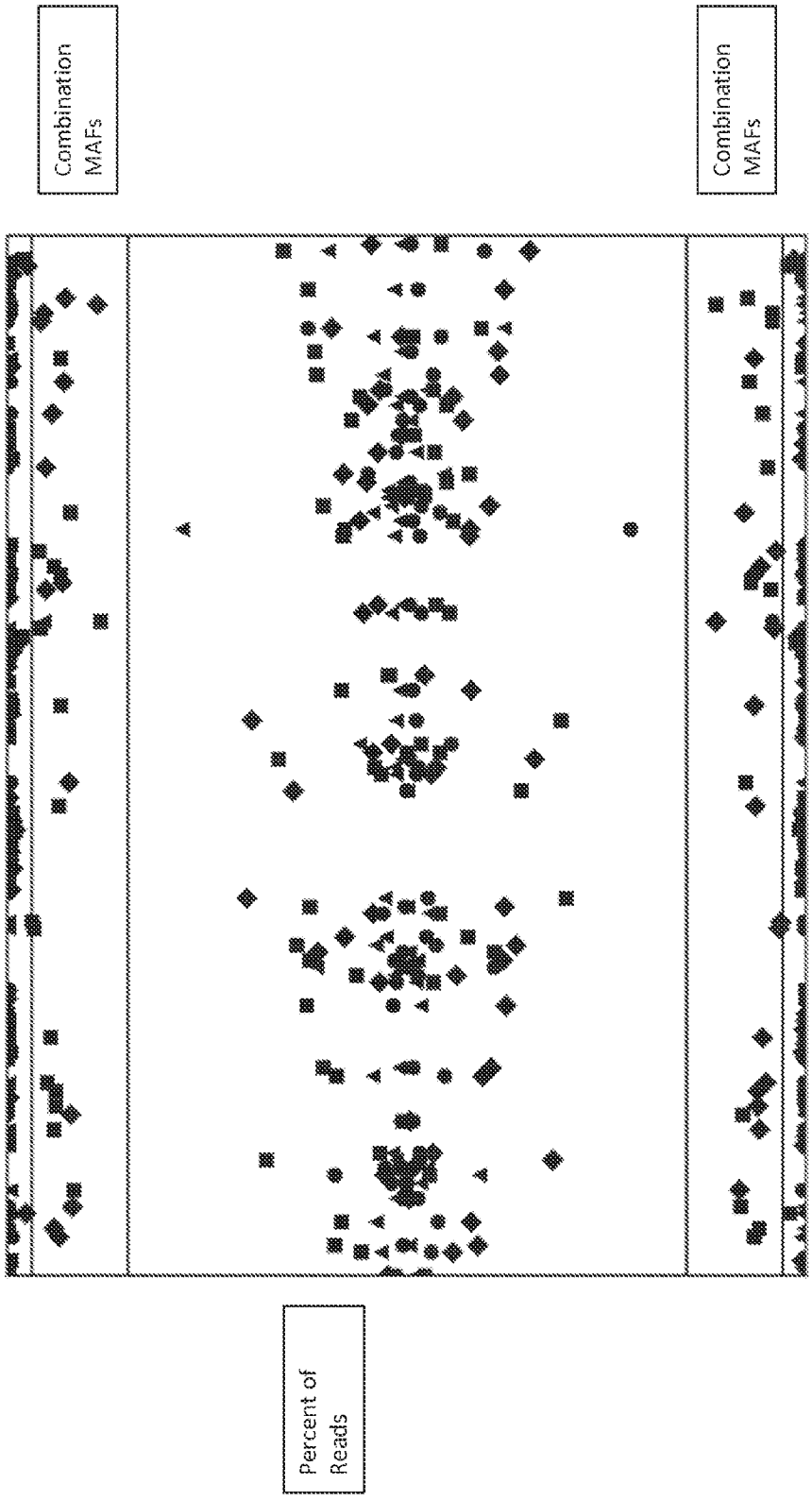


Figure 2

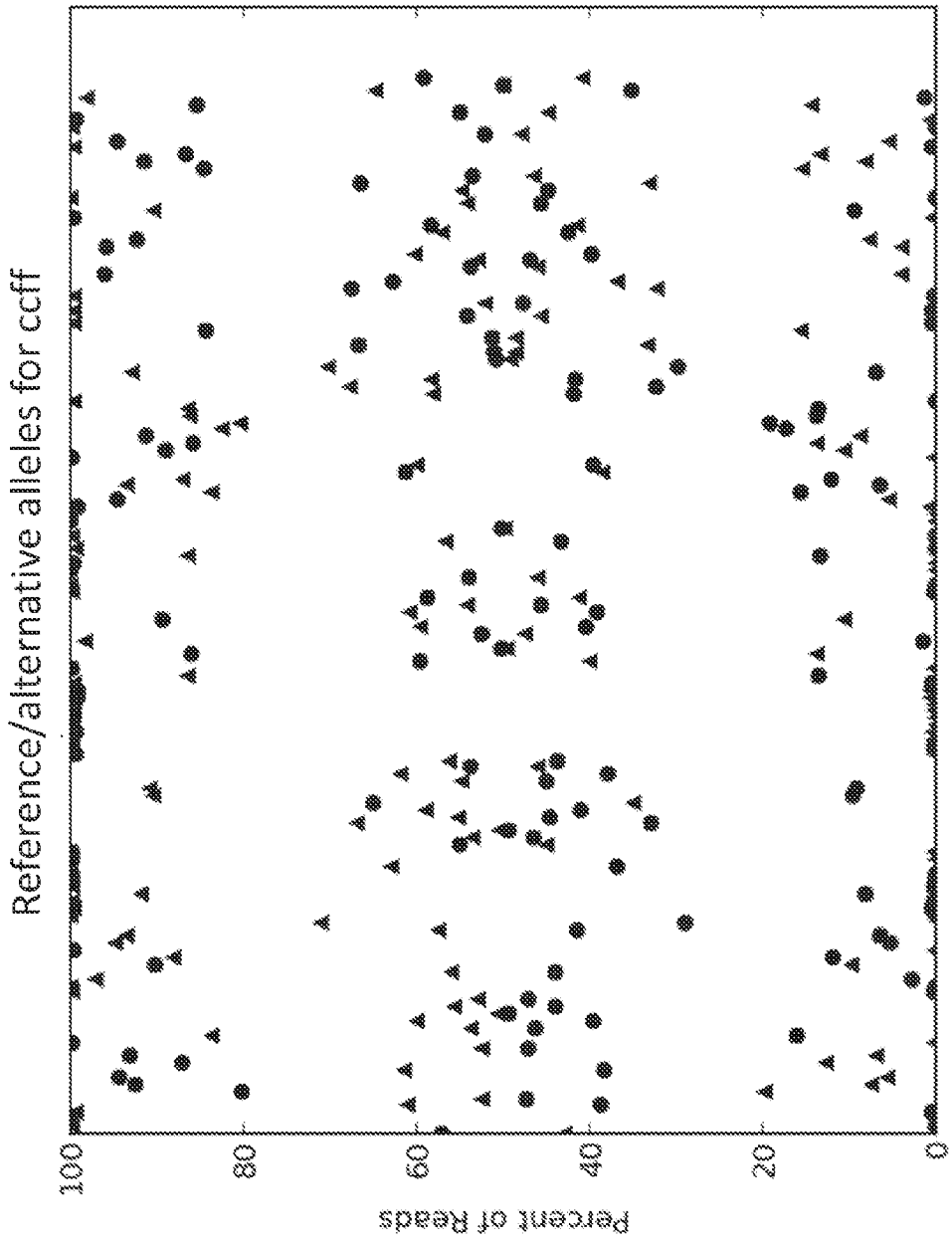

Reference/alternative alleles for ccff
Percent of Reads
100
80
60
40
20
0

Figure 3

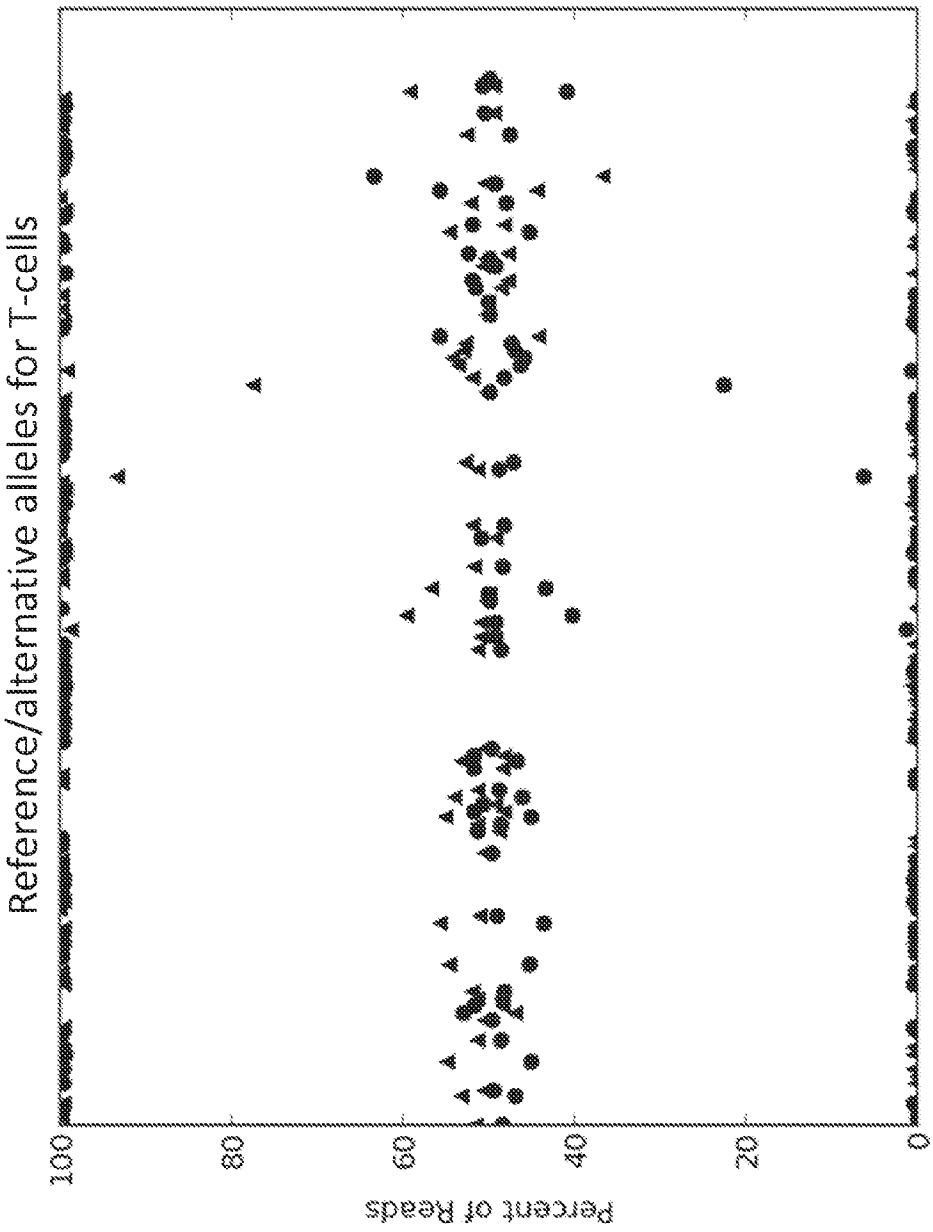

Reference/alternative alleles for T-cells
Percent of Reads
100
80
60
40
20
0

Figure 4

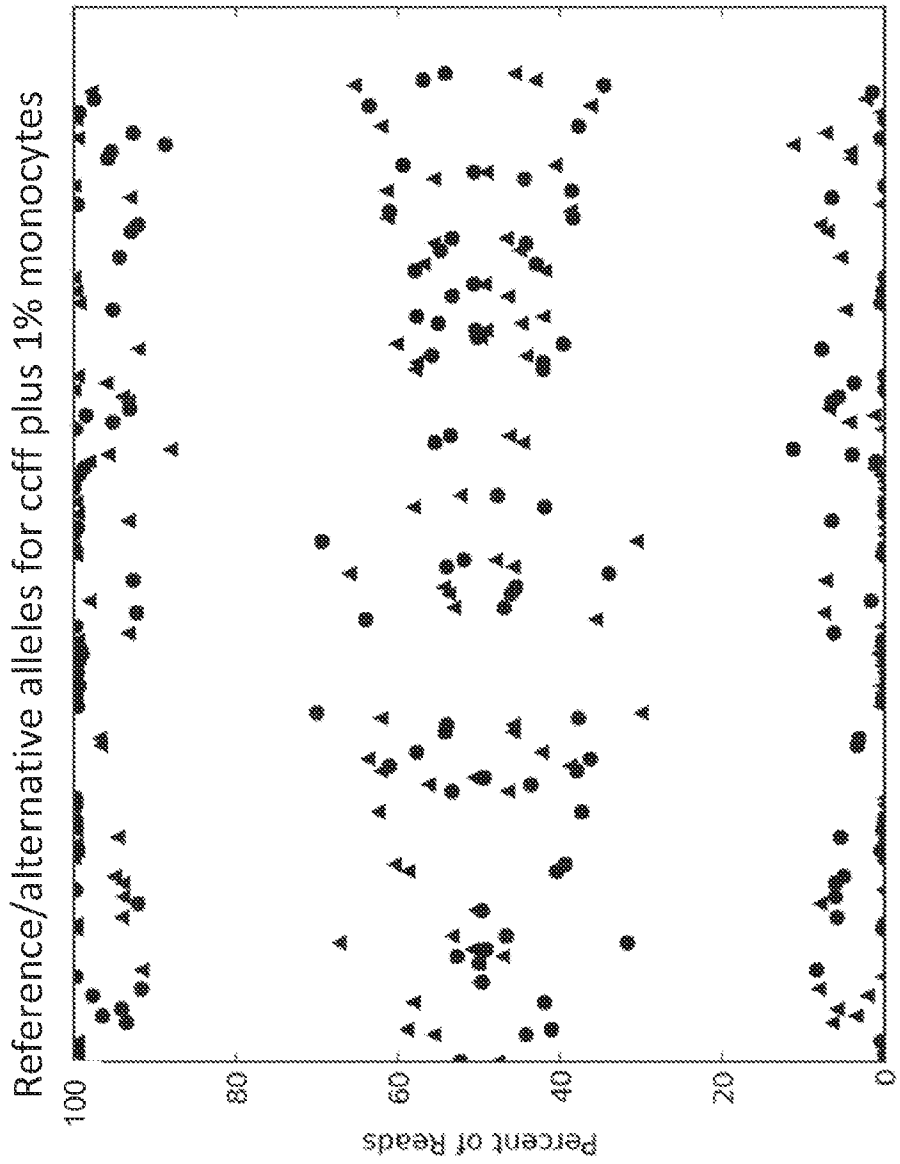

Reference/alternative alleles for ccff plus 1% monocytes
Percent of Reads
100
80
60
40
20
0

Figure 5

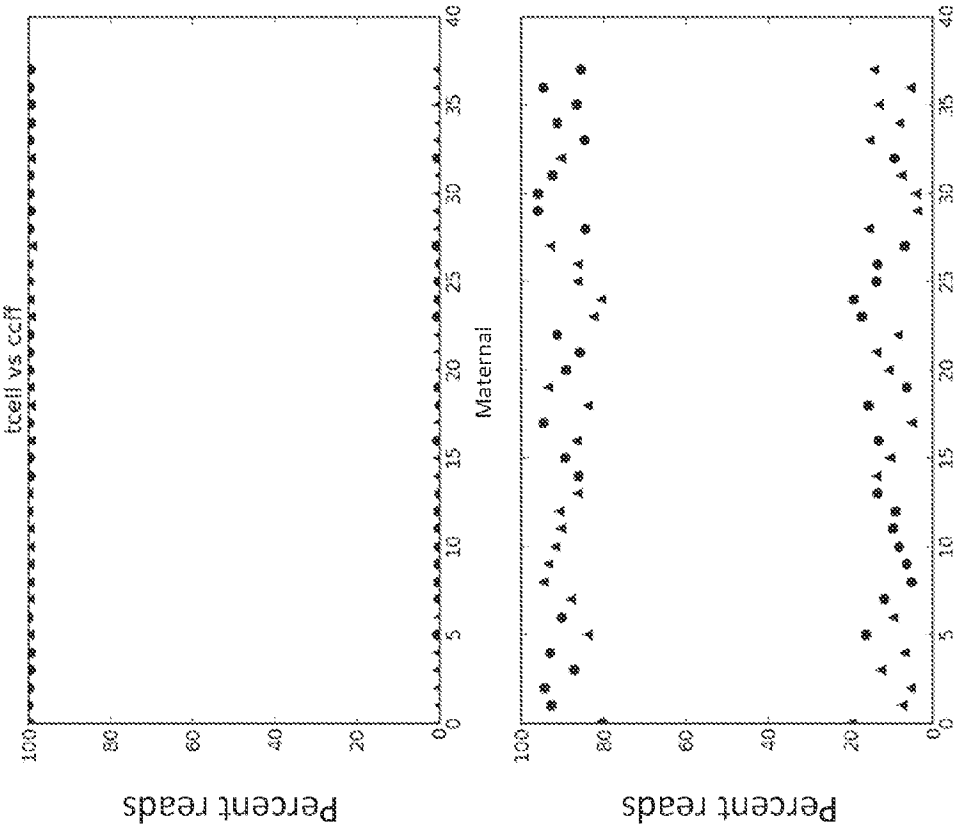

tcell vs ccff
Percent reads
Maternal
Percent reads
Combination

Figure 6

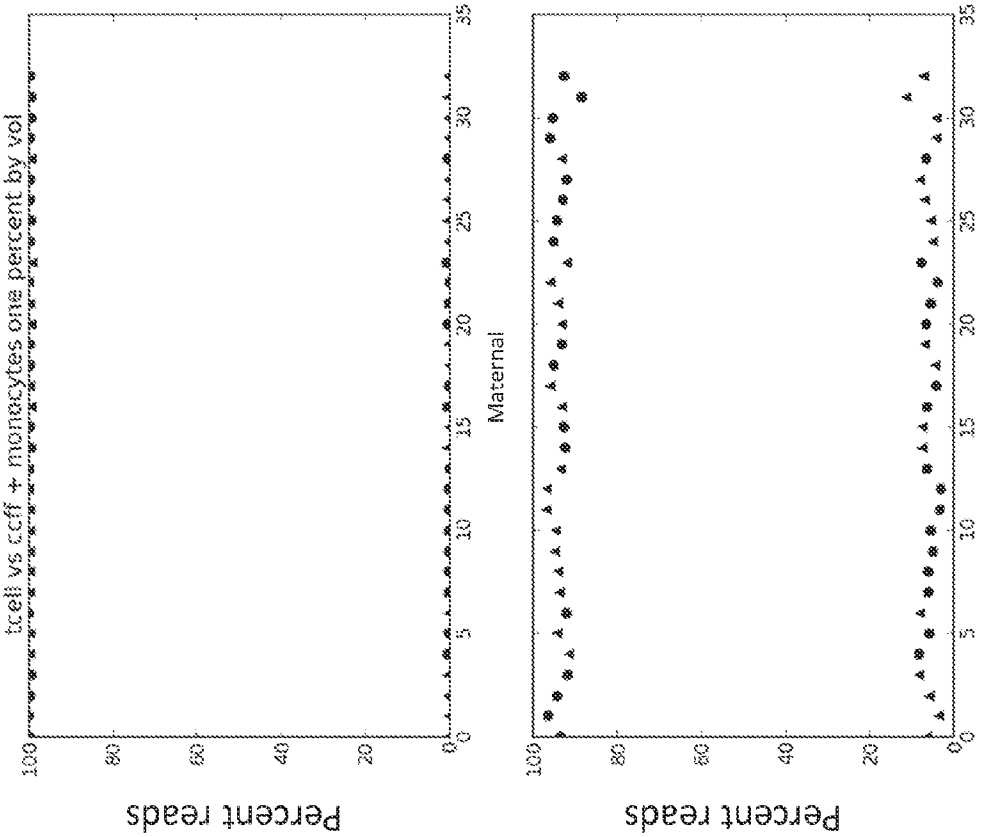

tcell vs ccff + monocytes one percent by vol
Percent reads
Maternal
Percent reads
Combination

Figure 7

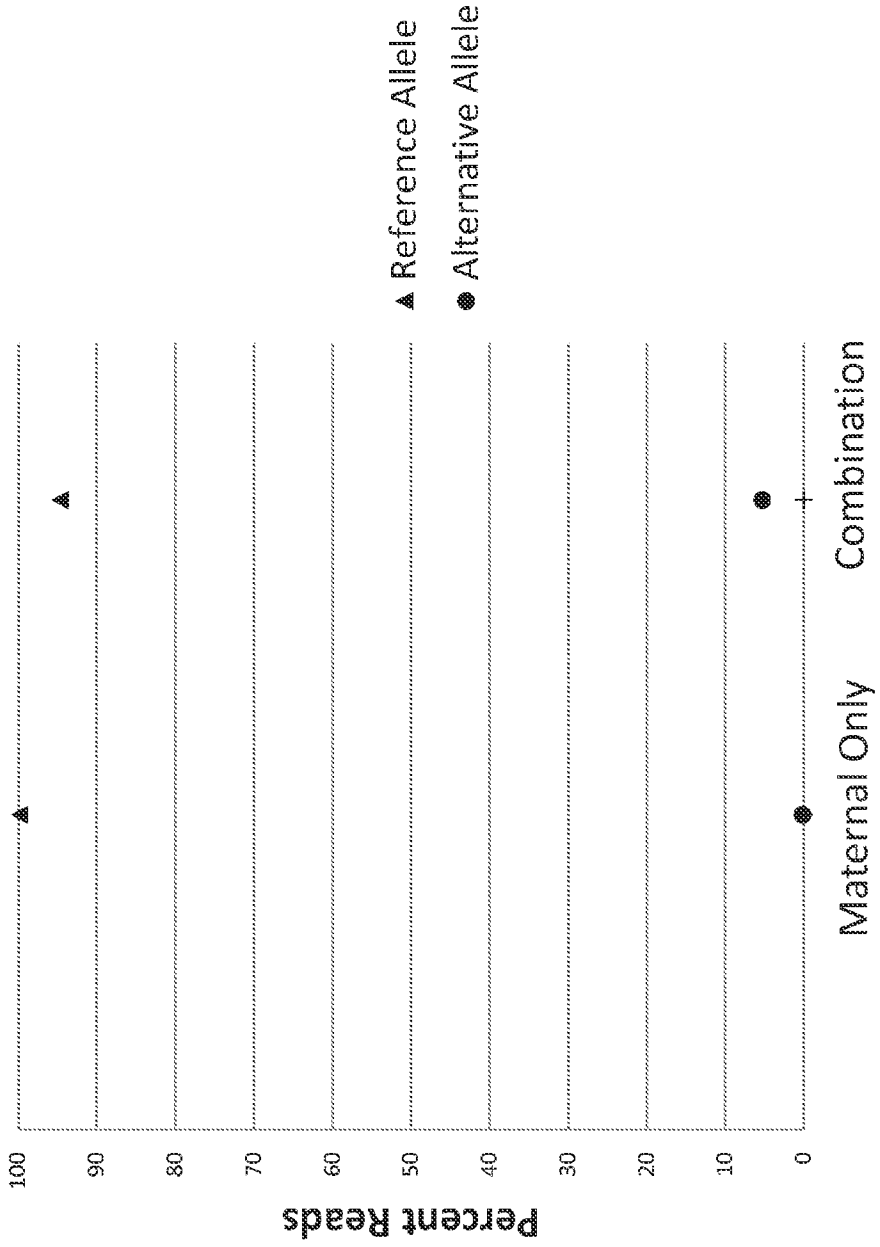

Reference Allele
Alternative Allele
Percent Reads
100
90
80
70
60
50
40
30
20
10
0
Maternal Only
Combination

Figure 8

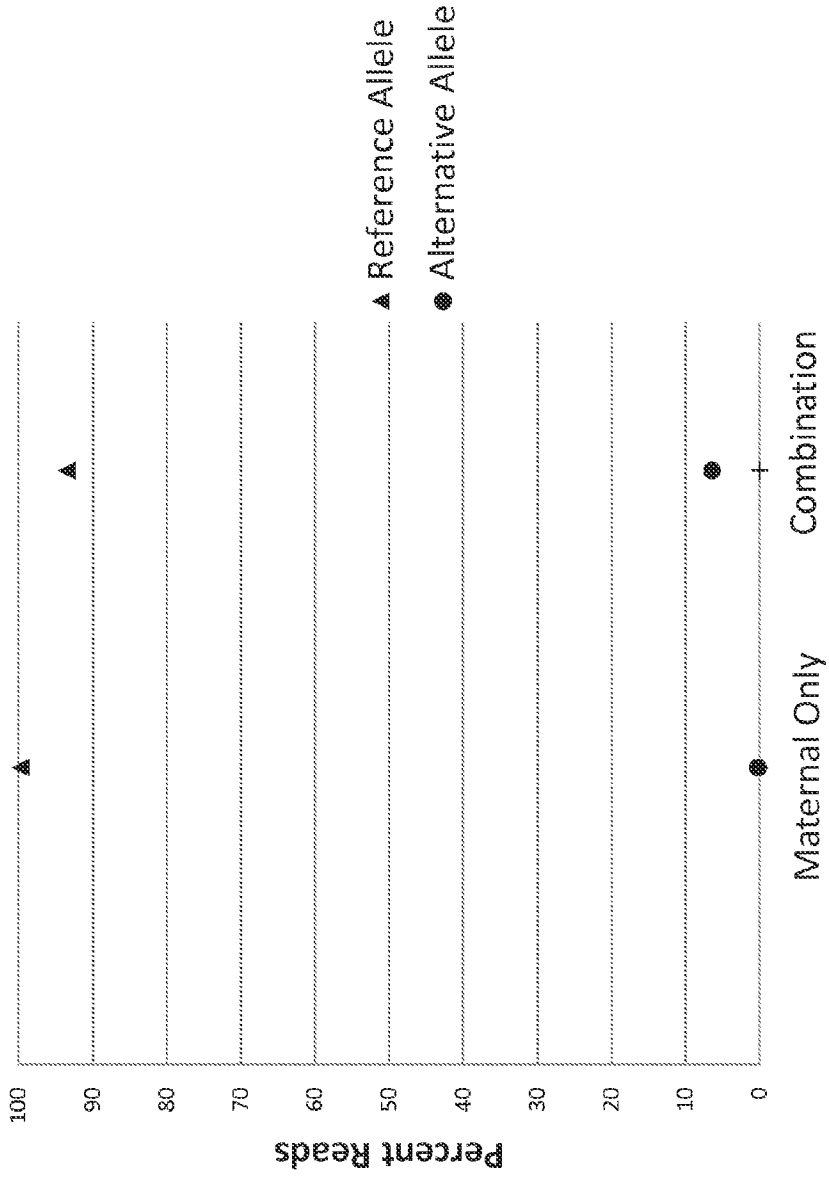

Reference Allele
Alternative Allele
Percent Reads
100
90
80
70
60
50
40
30
20
10
0
Maternal Only
Combination

Figure 9

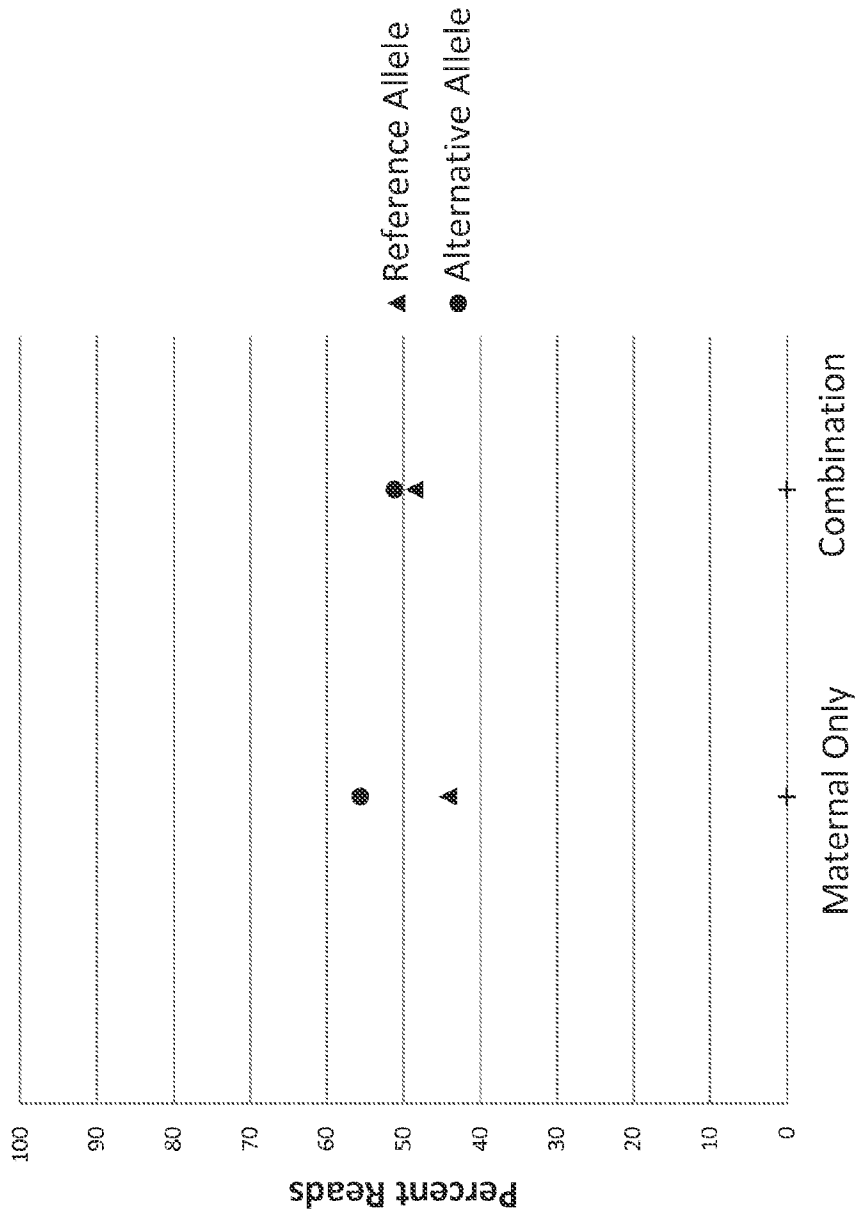

Reference Allele
Alternative Allele
Percent Reads
100
90
80
70
60
50
40
30
20
10
0
Maternal Only
Combination

Figure 10

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US13/46020

A. CLASSIFICATION OF SUBJECT MATTER
IPC(8) - C12Q 1/68; G01N 33/53; C40B 30/04 (2013.01)
USPC - 435/7.2; 506/9
According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
IPC(8): C12Q 1/68, 02; G01N 33/00, 53, 577; C40B 30/04 (2013.01)
USPC: 435/6.11, 6.1, 4, 7.24, 7.21, 7.2, 7.1, 7.92, 7.4, 29; 506/9

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)
MicroPatent (US-G, US-A, EP-A, EP-B, WO, JP-bib, DE-C,B, DE-A, DE-T, DE-U, GB-A, FR-A); Google; Google Scholar; DialogPRO; PubMed; diagnos*, prognos*, efficac*, profile, marker, biomarker, acellular, cell-free, bodily fluid, circulating, 'ctDNA,' 'cirDNA,' 'cfDNA,' phagocyte, neutrophil, monocyte, macrophage, 'mast cell,' 'dendritic cell,' microvesicle, lysing

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	US 2012/0053073 A1 (KASSIS, AI), March 1, 2012; abstract; paragraphs [0008]-[0018], [0027]-[0034], [0156]	1-24, 25/1-25/3, 25/10-25/12, 26/1-26/3, 26/10-26/12, 27/4-27/6, 27/13-27/15, 28/4-28/6, 28/13-28/15, 29/7-29/9, 29/16-29/18, 30/7-30/9, 30/16-30/18, 31/19-31/24, 32/19-32/24, 34/4-34/6, 34/13-34/15, 35/1-35/9, 35/19-35/21, 38/1-9, 38/19-21, 44/19-44/24, 71/1-71/3, 71/10-71/12, 71/19-71/24, 73/4-9, 73/13-18, 93-96, 147/6, 147/15

☒ Further documents are listed in the continuation of Box C. ☐

* Special categories of cited documents:
"A" document defining the general state of the art which is not considered to be of particular relevance
"E" earlier application or patent but published on or after the international filing date
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
"O" document referring to an oral disclosure, use, exhibition or other means
"P" document published prior to the international filing date but later than the priority date claimed
"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
05 November 2013 (05.11.2013)	12 NOV 2013

Name and mailing address of the ISA/US	Authorized officer:
Mail Stop PCT, Attn: ISA/US, Commissioner for Patents P.O. Box 1450, Alexandria, Virginia 22313-1450 Facsimile No. 571-273-3201	Shane Thomas PCT Helpdesk: 571-272-4300 PCT OSP: 571-272-7774

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

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US13/46020

C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	WO2012012693 A2 (KASSIS, AI), January 26, 2012; abstract; paragraphs [0007]-[0022], [0027]-[0034]	1-24, 25/1-25/3, 25/10-25/12, 26/1-26/3, 26/10-26/12, 27/4-27/6, 27/13-27/15, 28/4-28/6, 28/13-28/15, 29/7-29/9, 29/16-29/18, 30/7-30/9, 30/16-30/18, 31/19-31/24, 32/19-32/24, 34/4-34/6, 34/13-34/15, 35/1-35/9, 35/19-35/21, 38/1-9, 38/19-21, 44/19-44/24, 71/1-71/3, 71/10-71/12, 71/19-71/24, 73/4-9, 73/13-18, 94-96, 147/6, 147/15
A	WEILAND, M et al. Small RNAs Have A Large Impact: Circulating microRNAs As Biomarkers For Human Diseases. Epub. RNA Biology. June 1, 2012, Vol. 9, No. 6, pp 850-859.	1-9
A	CARPAGANO, GE et al. Microsatellite Alterations And Cell-Free DNA Analysis: Could They Increase the Cytology Sensitivity In The Diagnosis Of Malignant Pleural Effusion? Epub. Rejuvenation Research. 02 May 2012, Vol. 15, No. 3, pp 265-273.	1-9

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US13/46020

Box No. II Observations where certain claims were found unsearchable (Continuation of item 2 of first sheet)

This international search report has not been established in respect of certain claims under Article 17(2)(a) for the following reasons:

1. ☐ Claims Nos.:
because they relate to subject matter not required to be searched by this Authority, namely:

2. ☐ Claims Nos.:
because they relate to parts of the international application that do not comply with the prescribed requirements to such an extent that no meaningful international search can be carried out, specifically:

3. ☒ Claims Nos.: 33, 36, 37, 39-43, 45-70, 72, 74-92, 97-146, 148-150
because they are dependent claims and are not drafted in accordance with the second and third sentences of Rule 6.4(a).

Box No. III Observations where unity of invention is lacking (Continuation of item 3 of first sheet)

This International Searching Authority found multiple inventions in this international application, as follows:

1. ☐ As all required additional search fees were timely paid by the applicant, this international search report covers all searchable claims.

2. ☐ As all searchable claims could be searched without effort justifying additional fees, this Authority did not invite payment of additional fees.

3. ☐ As only some of the required additional search fees were timely paid by the applicant, this international search report covers only those claims for which fees were paid, specifically claims Nos.:

4. ☐ No required additional search fees were timely paid by the applicant. Consequently, this international search report is restricted to the invention first mentioned in the claims; it is covered by claims Nos.:

Remark on Protest

☐ The additional search fees were accompanied by the applicant's protest and, where applicable, the payment of a protest fee.

☐ The additional search fees were accompanied by the applicant's protest but the applicable protest fee was not paid within the time limit specified in the invitation.

☐ No protest accompanied the payment of additional search fees.

Espacenet

Bibliographic data: CN104053789 (A) — 2014-09-17

Method, System And Computer Readable Medium For Determining Base Information In Predetermined Area Of Fetus Genome

Inventor(s): CHEN SHENGPEI; GE HUIJUAN; LI XUCHAO; YI SHANG; WANG JIAN; WANG JUN; YANG HUANMING; ZHANG XIUQIN ± (CHEN SHENGPEI, ; GE HUIJUAN, ; LI XUCHAO, ; YI SHANG, ; WANG JIAN, ; WANG JUN, ; YANG HUANMING, ; ZHANG XIUQIN)

Applicant(s): BGI HEALTH SERVICE CO LTD ± (BGI HEALTH SERVICE CO LTD)

Classification: - international: ***C12Q1/68***
- cooperative: **C12Q1/6809 (CN); C12Q1/6858 (EP, KR, US); C12Q1/6869 (EP, KR, US); C12Q1/6883 (US); G16B20/00 (EP, KR, US); G16B30/00 (EP, KR, US); G16B40/00 (KR); G16H50/20 (EP, KR, US); G16H50/30 (EP, KR, US);** C12Q2600/118 (US); C12Q2600/156 (US) more

Application number: CN20128067404 20120514 Global Dossier

Priority number(s): WO2012CN75478 20120514

Also published as: CN104053789 (B) AU2012380221 (A1) AU2012380221 (B2) EP2851431 (A1) EP2851431 (A4) EP2851431 (B1) ES2656023 (T3) HK1196401 (A1) JP2015525062 (A) JP6045686 (B2) KR101770460 (B1) KR20140146193 (A) PL2851431 (T3) RU2014150655 (A) RU2597981 (C2) SG11201407515R (A) US2015094210 (A1) US2018320235 (A1) WO2013170429 (A1) less

Abstract of CN104053789 (A)

Provided are a method, system and computer readable medium for determining the base information in a predetermined area of a fetus genome, the method comprising the following steps: constructing a sequence library for the DNA samples of the fetus genome; sequencing the sequence library to obtain the sequencing result of the fetus, the sequencing result of the fetus comprised of a plurality of sequencing data; and based on the sequencing result of the fetus, determining the base information in the predetermined area according to the hidden Markov model in conjunction with the genetic information of an individual related hereditarily to the fetus.

(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 104053789 A
(43)申请公布日 2014.09.17

(21)申请号 201280067404.1
(22)申请日 2012.05.14
(85)PCT国际申请进入国家阶段日
2014.07.17
(86)PCT国际申请的申请数据
PCT/CN2012/075478 2012.05.14
(87)PCT国际申请的公布数据
WO2013/170429 ZH 2013.11.21
(71)申请人 深圳华大基因医学有限公司
地址 518083 广东省深圳市盐田区北山道146号北山工业区11栋2、3楼
(72)发明人 陈盛培 葛会娟 李旭超 易赏 汪建 王俊 杨焕明 张秀清
(74)专利代理机构 北京清亦华知识产权代理事务所(普通合伙) 11201
代理人 李志东
(51)Int.Cl.
C12Q 1/68(2006.01)

(54)发明名称

确定胎儿基因组中预定区域碱基信息的方法、系统和计算机可读介质

(57)摘要

提供了确定胎儿基因组中预定区域碱基信息的方法、系统和计算机可读介质。其中,确定胎儿基因组中预定区域碱基信息的方法,包括下列步骤:针对胎儿基因组DNA样本,构建测序文库;对测序文库进行测序,以便获得胎儿的测序结果,该胎儿的测序结果由多个测序数据构成;基于胎儿的测序结果,结合胎儿遗传相关个体的遗传信息,根据隐马尔可夫模型,确定预定区域的碱基信息。

CN 104053789 A

Espacenet

Bibliographic data: WO2014153757 (A1) — 2014-10-02

METHOD, SYSTEM, AND COMPUTER READABLE MEDIUM FOR DETERMINING BASE INFORMATION OF PREDETERMINED AREA IN FETAL GENOME

Inventor(s): YIN XUYANG [CN]; JIANG HUI [CN]; CHEN SHENGPEI [CN]; GONG CHUN [CN]; CHEN FANG [CN]; ZHANG CHUNLEI [CN]; PAN XIAOYU [CN] ± (YIN, XUYANG, ; JIANG, HUI, ; CHEN, SHENGPEI, ; GONG, CHUN, ; CHEN, FANG, ; ZHANG, CHUNLEI, ; PAN, XIAOYU)

Applicant(s): BGI SHENZHEN [CN] ± (BGI SHENZHEN)

Classification: - international: *C12Q1/68*
- cooperative: **C12Q1/6876 (EP); G16B20/00 (EP); G16B30/00 (EP);** C12Q2600/156 (EP)

Application number: WO2013CN73375 20130328 Global Dossier

Priority number(s): WO2013CN73375 20130328

Also published as: CN105051208 (A) CN105051208 (B) HK1213945 (A1)

Abstract of WO2014153757 (A1)

A method, a system, and a computer readable medium for determining base information of a predetermined area in a fetal genome are provided. The method for determining base information of a predetermined area in a fetal genome comprises the following steps: obtaining a sequencing result of a DNA sample of a genome of an embryo cell, and a sequencing result of a genome sample of an individual related to embryo heredity; constructing a heredity sketch of the embryo based on the sequencing result of the DNA sample of the genome of the embryo cell, to determine an initial genotype of the embryo; determining haplotypes of parents of the embryo based on the sequencing result of the genome sample of the individual related to the embryo heredity; and using the initial genotype of the embryo as an observation sequence according to the hidden markov model, and determining the base information of the predetermined area in the fetal genome based on the haplotypes of the parents of the embryo.

(12) 按照专利合作条约所公布的国际申请

(19) 世界知识产权组织
国 际 局

(43) 国际公布日
2014 年 10 月 2 日 (02.10.2014)

WIPO | PCT

(10) 国际公布号
WO 2014/153757 A1

(51) 国际专利分类号:
C12Q 1/68 (2006.01)

(21) 国际申请号: PCT/CN2013/073375

(22) 国际申请日: 2013 年 3 月 28 日 (28.03.2013)

(25) 申请语言: 中文

(26) 公布语言: 中文

(71) 申请人: **深圳华大基因研究院 (BGI SHENZHEN)** [CN/CN]; 中国广东省深圳市盐田区北山工业区综合楼, Guangdong 518083 (CN)。

(72) 发明人: **殷旭阳 (YIN, Xuyang)**; 中国广东省深圳市盐田区北山工业区综合楼, Guangdong 518083 (CN)。 **蒋慧 (JIANG, Hui)**; 中国广东省深圳市盐田区北山工业区综合楼, Guangdong 518083 (CN)。 **陈盛培 (CHEN, Shengpei)**; 中国广东省深圳市盐田区北山工业区综合楼, Guangdong 518083 (CN)。 **龚淳 (GONG, Chun)**; 中国广东省深圳市盐田区北山工业区综合楼, Guangdong 518083 (CN)。 **陈芳 (CHEN, Fang)**; 中国广东省深圳市盐田区北山工业区综合楼, Guangdong 518083 (CN)。 **张春雷 (ZHANG, Chunlei)**; 中国广东省深圳市盐田区北山工业区综合楼, Guangdong 518083 (CN)。 **潘小瑜 (PAN, Xiaoyu)**; 中国广东省深圳市盐田区北山工业区综合楼, Guangdong 518083 (CN)。

(74) 代理人: **北京清亦华知识产权代理事务所（普通合伙） (TSINGYIHUA INTELLECTUAL PROPERTY LLC)**; 中国北京市海淀区清华园清华大学照澜院商业楼 301 室, Beijing 100084 (CN)。

(81) **指定国** (除另有指明，要求每一种可提供的国家保护): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, 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, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW。

(84) **指定国** (除另有指明，要求每一种可提供的地区保护): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, SZ, TZ, UG, ZM, ZW), 欧亚 (AM, AZ, BY, KG, KZ, RU, TJ, TM), 欧洲 (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG)。

本国际公布:

— 包括国际检索报告(条约第 21 条(3))。

(54) **Title**: METHOD, SYSTEM, AND COMPUTER READABLE MEDIUM FOR DETERMINING BASE INFORMATION OF PREDETERMINED AREA IN FETAL GENOME

(54) **发明名称** : 确定胚胎基因组中预定区域碱基信息的方法、系统和计算机可读介质

(57) **Abstract**: A method, a system, and a computer readable medium for determining base information of a predetermined area in a fetal genome are provided. The method for determining base information of a predetermined area in a fetal genome comprises the following steps: obtaining a sequencing result of a DNA sample of a genome of an embryo cell, and a sequencing result of a genome sample of an individual related to embryo heredity; constructing a heredity sketch of the embryo based on the sequencing result of the DNA sample of the genome of the embryo cell, to determine an initial genotype of the embryo; determining haplotypes of parents of the embryo based on the sequencing result of the genome sample of the individual related to the embryo heredity; and using the initial genotype of the embryo as an observation sequence according to the hidden markov model, and determining the base information of the predetermined area in the fetal genome based on the haplotypes of the parents of the embryo.

(57) **摘要**: 提供了确定胚胎基因组中预定区域碱基信息的方法、系统和计算机可读介质。其中，确定胚胎基因组中预定区域碱基信息的方法，包括下列步骤：获得胚胎细胞基因组 DNA 样本的测序结果，以及胚胎遗传相关个体基因组样本的测序结果；基于胚胎细胞基因组 DNA 样本的测序结果，构建该胚胎的遗传草图，以便确定胚胎初始基因型；基于胚胎遗传相关个体基因组样本的测序结果，确定胚胎父母的单倍体型；以及根据隐马尔可夫模型，以该胚胎初始基因型作为观察序列，基于该胚胎父母的单倍体型，确定胚胎基因组中预定区域的碱基信息。

WO 2014/153757 A1

确定胚胎基因组中预定区域碱基信息的方法、系统和计算机可读介质

优先权信息

无

技术领域

本发明涉及确定胚胎基因组中预定区域碱基信息的方法、系统和计算机可读介质。

背景技术

遗传性疾病是由于遗传物质发生改变而造成的疾病，具有先天性、家族性、终身性和遗传性的特点。遗传性疾病可分为 3 个大类：单基因遗传病、多基因遗传病及染色体异常。其中单基因病多由于单个致病基因的显性或隐性遗传所致基因功能异常；而多基因遗传病则是由多个基因变化影响所致的疾病，会在一定程度上受到外界环境因素的影响；染色体异常包括数目异常和结构异常，最为多见的是由于第 21 号染色体三体所致的唐氏综合症，患儿表现为先天愚型和肢体形状异常等其他先天性特征。由于目前对遗传性疾病尚无有效的治疗方式，只能针对性地进行支持治疗或者药物缓解，费用昂贵，给社会和家庭带来沉重经济和精神负担。因此，在孩子出生前就对孩子的患病状态进行检测，做好预防工作，以达到优生优育的目的，是十分必要的。例如在体外生殖过程中，在胚胎植入子宫之前，对胚胎的患病状态进行检测，将具有重大意义。对配子或植入到子宫腔之前的胚胎进行遗传学分析，将排除致病因素的正常胚胎植入子宫，从而防止患有遗传病的孩子的妊娠。对具有高风险妊娠遗传缺陷胚胎的夫妇实施植入前胚胎的患病状态的检测，无疑能有效降低患病孩子的出生，并且避免选择性终止妊娠给孕妇及家庭带来的诸多危害。

然而，目前的相关检测手段仍有待改进。

发明内容

本发明是基于发明人的下列发现而完成的：

发明人发现，植入前胚胎中可以活检取样进行检测的细胞数有限。例如培养至第三天的卵裂球仅有 4-8 个细胞，可以进行活检取样的仅 1-2 个细胞。即使在培养至第五天的囊胚，进行外滋养层细胞活检取样也只能取到 3-10 个细胞。由于单个或少数胚胎细胞的 DNA 量有限，难以直接进行全面的遗传变异检测，必需经过全基因组扩增(Whole Genome Amplification, WGA)，达到足够 DNA 量，才能对遗传变异做全面分析。而全基因组扩增常存在偏好性，会产生一定的杂合缺失或等位基因脱扣(Allele Drop Out)等现象，这些都为胚胎细胞的遗传病相关的遗传变异检测带来风险。

因此，很有必要开发能够对扩增引起的等位基因脱扣等错误进行定位和矫正的方法，以提高遗传变异检测的准确性。

本发明旨在至少解决现有技术中存在的技术问题之一。通过本发明方法推算胚胎单

倍体型，可以对发生等位基因脱扣的位点进行定位，并能够进行矫正。而胚胎单倍体型分析方法可以基于家系成员如父母和先证者的遗传信息，或者基于来自同一父母的其他胚胎细胞的遗传信息进行推算。

在本发明的一个方面，本发明提出了一种确定胚胎基因组中预定区域碱基信息的方法。该方法结合系谱推断和统计算法来进行胚胎单体型的确定，通过家系中相关个体的基因型，即追溯染色体片段的传递，结合统计算法来推断单体型状态。根据本发明的实施例，该方法包括下列步骤：获得胚胎细胞基因组 DNA 样本的测序结果，以及胚胎遗传相关个体基因组样本的测序结果；基于胚胎细胞基因组 DNA 样本的测序结果，构建所述胚胎的遗传草图，以便确定胚胎初始基因型；基于所述胚胎遗传相关个体基因组样本的测序结果，确定胚胎父母的单倍体型；以及根据隐马尔可夫模型，以所述胚胎初始基因型作为观察序列，基于所述胚胎父母的单倍体型，确定胚胎基因组中预定区域的碱基信息。

需要说明的是，子代的基因组形成过程，相当于亲代基因组的一次随机重组（即连锁互换单倍体型重组，以及配子的随机组合）。将胚胎的单倍型作为隐含状态（hidden states），将胚胎单细胞全基因组扩增后的测序数据作为观察序列（observations），利用贝叶斯算法借助先验数据推算出状态转移概率（transition probabilities），构建观察序列概率分布（observation symbol probabilities）和初始状态概率分布（initial state distribution），然后通过惠特比算法（Viterbi algorithm）可以推断出最可能的胚胎单体型组合。因而，根据本发明的实施例，借助隐马尔可夫模型，例如可以通过利用惠特比算法（Viterbi algorithm），参考胚胎遗传相关个体的遗传信息，可以确定胚胎基因组中特定区域的核酸序列，由此，可以有效地对胚胎基因组的遗传信息进行植入前检测。对于在胚胎细胞全基因组扩增 DNA 测序结果中覆盖程度较低，存在杂合缺失或等位基因脱扣(Allele Drop Out)的位点或 DNA 序列，可以通过本方法准确地推断得到。因此通过本方法，可以对杂合缺失或等位基因脱扣(Allele Drop Out)的位点或序列进行校正，使检测结果更加准确可靠。

在本发明的又一方面，本发明提出了一种用于确定胚胎基因组中预定区域碱基信息的系统。根据本发明的实施例，该系统包括：文库构建装置，所述文库构建装置适于针对胚胎基因组 DNA 样本以及胚胎遗传相关个体的基因组 DNA 样本，分别构建测序文库；测序装置，所述测序装置与所述文库构建装置相连，并且适于对所述测序文库进行测序，以便获得胚胎以及胚胎遗传相关个体的测序结果；以及分析装置，所述分析装置与所述测序装置相连，并且适于：基于胚胎基因组 DNA 样本的测序结果，构建所述胚胎的遗传草图，以便确定胚胎初始基因型；基于所述胚胎遗传相关个体基因组样本的测序结果，确定胚胎父母的单倍体型；以及根据隐马尔可夫模型，以所述胚胎初始基因型作为观察序列，基于所述胚胎父母的单倍体型，确定胚胎基因组中预定区域的碱基信息。根据本发明的实施例，还可以进一步包括胚胎活检装置，所述胚胎活检装置适于针对体外培养的胚胎进行活检，对胚胎细胞进行取样。利用该系统，能够有效地实施前面所述的确定胚胎基因组中预定区域碱基信息的方法，可以借助隐马尔可夫模型，例如可以通过利用惠特比算法（Viterbi algorithm），参考胚胎遗传相关个体的遗传信息，可以确定胚胎基因组中特定区域的核酸序

列，由此，可以有效地对胚胎基因组的遗传信息进行植入前检测，从而可以有效地对胚胎基因组的遗传信息进行确定。

在本发明的另一方面，本发明还提出了一种计算机可读介质。根据本发明的实施例，基于胚胎细胞基因组DNA样本的测序结果，构建胚胎的遗传草图，以便确定胚胎初始基因型；基于所述胚胎遗传相关个体基因组样本的测序结果，确定胚胎父母的单倍体型；以及以根据隐马尔可夫模型，所述胚胎初始基因型作为观察序列，基于所述胚胎父母的单倍体型，确定胚胎基因组中预定区域的碱基信息。利用本发明的计算机可读介质，能够有效地被处理器执行其存储的指令，以便借助隐马尔可夫模型，例如可以通过利用惠特比算法（Viterbi algorithm），基于胚胎细胞的测序结果，参考胚胎遗传相关个体的遗传信息，可以确定胚胎基因组中特定区域的核酸序列，由此，可以有效地对胚胎基因组的遗传信息进行植入前检测。

本发明的附加方面和优点将在下面的描述中部分给出，部分将从下面的描述中变得明显，或通过本发明的实践了解到。

附图说明

本发明的上述和/或附加的方面和优点从结合下面附图对实施例的描述中将变得明显和容易理解，其中：

图1为根据本发明一个实施例的利用隐马尔可夫模型进行分析的流程示意图；以及

图2为根据本发明的一个实施例的用于确定胚胎基因组中预定区域核酸序列的系统的结构示意图。

发明详细描述

下面详细描述本发明的实施例，所述实施例的示例在附图中示出，其中自始至终相同或类似的标号表示相同或类似的元件或具有相同或类似功能的元件。下面通过参考附图描述的实施例是示例性的，仅用于解释本发明，而不能理解为对本发明的限制。

需要说明的是，术语“第一”、“第二”仅用于描述目的，而不能理解为指示或暗示相对重要性或者隐含指明所指示的技术特征的数量。由此，限定有“第一”、“第二”的特征可以明示或者隐含地包括一个或者更多个该特征。进一步地，在本发明的描述中，除非另有说明，“多个”的含义是两个或两个以上。

确定胚胎基因组中预定区域碱基信息的方法

在本发明的第一方面，本发明提出了一种确定胚胎基因组中预定区域碱基信息的方法。根据本发明的实施例，该方法包括下列步骤：

首先，获得胚胎细胞基因组DNA样本的测序结果，以及胚胎遗传相关个体基因组样本的测序结果。其中，在本文中所使用的术语“胚胎遗传相关个体”指的是在遗传意义上，与胚胎之间具有亲缘关系的个体，例如根据本发明的实施例，可以采用的“胚胎遗传相关个体”为胚胎的亲代例如父母、祖父母、外祖父母以及胚胎父母的其他子女。这里所说的

胚胎父母的其他子女应做广义理解，既可以是已经出生的子女，也可以是尚未出生的子女（受精卵或胚胎），也可以是死亡的胚胎，也可以是体外培养的胚胎或受精卵，只要与待检测胚胎共有相同的父母即可。

根据本发明的实施例，胚胎细胞基因组 DNA 样本的来源并不受特别限制。根据本发明的实施例，具体地，可以针对胚胎进行活检取样，获得胚胎细胞，并对胚胎细胞样本进行全基因组扩增(WGA)，以便获得胚胎细胞基因组 DNA。在本文中所采用的术语“胚胎活检”是指利用显微技术，从胚胎中分离部分胚胎细胞，或从受精卵/配子中分离部分细胞的技术。其中，根据本发明的实施例，该胚胎细胞可来自于卵裂球、囊胚滋养层、受精卵和配子中的任一种，可以是单个细胞，也可以是 2-10 个细胞。此外，也可以采用任何含有胚胎核酸的孕妇样本进行全基因组扩增，以便获得胚胎基因组 DNA 样本。由此，可以在不影响胚胎正常发育的前提下，对胚胎的基因组进行有效监测。其中，“全基因组扩增(WGA)”主要包括多重链置换扩增(Multiple Displacement Amplification, MDA)和基于 PCR 的 WGA 技术，可以采用自主开发的 WGA 扩增流程，也可以采用商业化的试剂盒如 Qiagen 公司的 REPLI-g 系列的试剂盒、Sigma-Aldrich 公司的 GenomePlex WGA 试剂盒(WGA4)、Rubicon Genomics 公司的 PicoPlex WGA 试剂盒、GE-Healthcare 公司的 illustra Genomiphi WGA 试剂盒等。在扩增 DNA 样本之后，针对胚胎细胞的扩增 DNA 样本以及胚胎遗传相关个体的基因组 DNA 样本，分别构建测序文库。

关于针对核酸样本，构建测序文库的方法和流程，本领域技术人员可以根据不同的测序技术进行适当选择，关于流程的细节，可以参见测序仪器的厂商例如 Illumina 公司所提供的规程，例如参见 Illumina 公司 Multiplexing Sample Preparation Guide （Part#1005361; Feb 2010）或 Paired-End Sample Prep Guide（Part#1005063; Feb 2010），通过参照将其并入本文。根据本发明的实施例，从生物样本提取核酸样本的方法和设备，也不受特别限制，可以采用商品化的核酸提取试剂盒进行。

在构建测序文库后，将测序文库应用于测序仪器，对测序文库进行测序，并获得相应的测序结果，该测序结果是由多个测序数据构成的。根据本发明的实施例，可以用于进行测序的方法和设备并不受特别限制，包括但不限于双脱氧链终止法；优选高通量的测序方法，由此，能够利用这些测序装置的高通量、深度测序的特点，进一步提高测序效率。从而，能够提高后续对测序数据进行分析，尤其是统计检验分析时的精确性和准确度。所述高通量的测序方法包括但不限于第二代测序技术或者是单分子测序技术。所述第二代测序平台（Metzker ML. Sequencing technologies-the next generation. Nat Rev Genet. 2010 Jan;11(1):31-46）包括但不限于 Illumina-Solexa（GA™,HiSeq2000™, Miseq 等）、ABI-SOLiD、Life Technologies-Ion Torrent/Proton 和 Roche-454（焦磷酸测序）测序平台；单分子测序平台（技术）包括但不限于 Helicos 公司的真实单分子测序技术（True Single Molecule DNA sequencing），Pacific Biosciences 公司单分子实时测序（single molecule real-time (SMRT™)），以及 Oxford Nanopore Technologies 公司的纳米孔测序技术等（Rusk, Nicole (2009-04-01). Cheap Third-Generation Sequencing. Nature Methods 6 (4): 244-245）。随着测序

技术的不断进化，本领域技术人员能够理解的是还可以采用其他的测序方法和装置进行全基因组测序。根据本发明的具体示例，可以利用选自 Illumina-Solexa、ABI-SOLiD、Life Technologies-Ion Torrent/Proton、Roche-454 和单分子测序装置的至少一种对所述全基因组测序文库进行测序。

根据本发明的实施例，在得到测序结果之后，可以基于胚胎细胞基因组 DNA 样本的测序结果，构建所述胚胎的遗传草图，以便确定胚胎初始基因型。基于所述胚胎遗传相关个体基因组样本的测序结果，确定胚胎父母的单倍体型。根据本发明的实施例，通过将胚胎细胞基因组 DNA 样本的测序结果与参考序列进行比对，构建所述胚胎的遗传草图。通过将所述胚胎遗传相关个体基因组样本的测序结果与参考序列进行比对，确定所述胚胎遗传相关个体的基因型；以及基于所述胚胎遗传相关个体的基因型，确定所述胚胎父母的单倍体型。根据本发明的实施例，可以采用已知的人类参考基因组作为参考序列。例如，根据本发明的实施例，采用的人类参考基因组为 NCBI 36.3，HG18。另外，根据本发明的实施例，进行比对的方法并不受特别限制。根据本发明的具体实施例，可以采用 SOAP 进行比对。

最后，以所述胚胎初始基因型作为观察序列，基于所述胚胎父母的单倍体型，根据隐马尔可夫模型，确定胚胎基因组中预定区域的碱基信息。

在本文中所使用的术语“预定区域”应作广义理解，是指任何包含可能发生预定事件位点的核酸分子的区域。对于 SNP 分析而言，可以是指包含 SNP 位点的区域。对于分析染色体非整倍性，则预定区域指的是所要分析的染色体的全长或者部分，即选择所有来自该染色体的测序数据。从测序结果中选择来自相应区域的测序数据的方法可以不受特别限制。根据本发明的实施例，可以通过将所得到的所有测序数据与已知的核酸参考序列进行比对，从而得到来自于预定区域的测序数据。另外，根据本发明的实施例，预定区域也可以是基因组上不连续的多个分散点。根据本发明的实施例，可以使用的参考序列的类型并不受特别限制，可以为任何含有感兴趣区域的已知序列。

根据本发明的实施例，可以通过基于所述胚胎细胞的测序结果，结合胚胎遗传相关个体的遗传信息，根据隐马尔可夫模型，确定所述预定区域的碱基信息。根据本发明的实施例，可以借助隐马尔可夫模型，利用惠特比算法（Viterbi algorithm），确定胚胎基因组中特定区域的碱基信息。由此，可以有效地对胚胎基因组的遗传信息进行植入前检测。

下面参考图 1，对借助隐马尔可夫模型利用惠特比算法进行分析的原理进行详细描述：

如前所述，在本文中所使用的术语“胚胎遗传相关个体”指的是在遗传意义上，与胚胎之间具有亲缘关系的个体，例如根据本发明的实施例，可以采用的“胚胎遗传相关个体”为胚胎的亲代例如父母、祖父母、外祖父母以及胚胎父母的其他子女。这里所说的胚胎父母的其他子女应做广义理解，既可以是已经出生的子女，也可以是尚未出生的子女（胚胎或受精卵），也可以是死亡的胚胎，也可以是体外培养的胚胎或受精卵，只要与待检测胚胎共有相同的父母即可。

由此，子代的基因组形成过程，相当于亲代基因组的一次随机重组（即连锁互换单倍体型重组，以及配子的随机组合）。将胚胎的单倍型作为隐含状态（hidden states），将胚胎

单细胞全基因组扩增后的测序数据作为观察序列（observations），借助先验数据推算出状态转移概率（transition probabilities），构建确定观察序列概率分布（observation symbol probabilities）和初始状态概率分布（initial state distribution），然后通过惠特比算法（Viterbi algorithm）可以推断出最可能的胚胎单体型组合。因而，根据本发明的实施例，借助隐马尔可夫模型，例如可以通过利用惠特比算法（Viterbi algorithm），参考胚胎遗传相关个体的遗传信息，可以确定胚胎基因组中特定区域的核酸序列，由此，可以有效地对胚胎基因组的遗传信息进行植入前检测。

详细分析步骤如下:

标记:

I. 需要检测的位点数为 N。

II. 观测序列;

III. 隐藏状态集为: $S=\{0,1\}$，定义了父母哪条染色体遗传给了子代。0 代表遗传给了先证者的那条染色体，1 代表没有遗传给先证者的那条染色体。

IV. 观测状态集为: $V=\{0,1\}$，1 代表遗传给先证者的染色体与胚胎遗传草图基因型一致，0 代表不一致。

第一步，构建初始状态概率分布向量，单倍体重组转移矩阵以及观察序列概率矩阵

I. 初始状态概率分布记为 $\pi=\{\pi_i\},(i\in S,\pi_i=0.5)$,

在没有参考数据的情况下，可以假设每种隐藏状态出现的可能性相等。

II. 记单倍体重组转移矩阵记为 $A=\{a_{ij}\},(i,j\in S)$，即隐藏状态转移的概率。

$$a_{ij}=\begin{cases}1-Nr/Np & i=j\\ Nr/Np & i\neq j\end{cases},$$

Nr、Np 分别表示期望重组数和单核苷酸多态性位点数，根据先验数据 Nr 可取 20-40 之间的自然数，用来计算隐藏状态的转移概率，即计算每个碱基组合的单体型发生重组的概率。

III. 记观察序列概率矩阵为 $B=\{b_i(k)\},(i\in S,k\in V)$，这里的概率分布 $b_i(k)$ 用子代一定是纯合（Must-hom）或一定是杂合(Must-het)的位点来估算。

$$b_i(0)=\frac{\#sites(L>0,\text{Must-hom.})}{\#sites(L>0,\text{Must-het. or Must-hom.})},\quad b_i(1)=1-b_i(0)。$$

第二步，构建局部概率矩阵和逆向指针

定义 局部概率：$\delta_t(j)=\max\left[\delta_{t-1}(i)\cdot a_{ij}\right]\cdot b_j(K)\quad t\in\{1...N\}$。

定义 逆向指针：$\Psi_i(j)=\arg\max \delta_{t-1}(i)\cdot a_{ij} \quad t\in\{1...N\}$。

第三步，确定最终状态状态，并回溯最优路径

确定最终状态，$q_N^*=\arg\max_{q_N\in S}\delta_N(q_N)$。

回溯最优路径，即最可能胚胎单基因型 $q_t^*=\Psi_{t+1}(q^*_{t+1})$ $(t=1,2,3,...,N-1)$。

第四步，输出结果

即，根据本发明的实施例，根据隐马尔可夫模型，确定胚胎基因组中预定区域的碱基信息进一步包括:

构建初始状态概率分布向量、隐藏状态转移的概率矩阵以及观察序列概率矩阵;

利用惠特比算法确定最终状态并回溯最优路径，以便确定胚胎基因组中预定区域的碱基信息。

根据具体的实施例，所述隐马尔可夫模型采用下列参数:

初始状态概率分布为 $\pi=\{\pi_i\},(i\in S,\pi_i=0.5)$,

隐藏状态转移的概率矩阵为 $A=\{a_{ij}\},(i,j\in S)$，其中，$a_{ij}=\begin{cases}1-Nr/Np & i=j\\ Nr/Np & i\neq j\end{cases}$

Nr、Np 分别表示期望重组数和单核苷酸多态性位点数，Nr 为自然数，取值范围 20-40，

观察序列概率矩阵为 $B=\{b_i(k)\},(i\in S,k\in V)$，其中，$b_i(0)=\frac{\#sites(L>0,\text{Must-hom.})}{\#sites(L>0,\text{Must-het. or Must-hom.})}$, $b_i(1)=1-b_i(0)$,

#sites (L>0,Must-hom.)为子代一定为纯合的位点的数目，#sites (L>0,Must-hom. or Must-het.)为子代一定为纯合的位点的数目以及子代一定为杂合的位点的数目的总和;

局部概率为 $\delta_t(j)=\max\left[\delta_{t-1}(i)\cdot a_{ij}\right]\cdot b_j(K) \quad t\in\{1...N\}$,

逆向指针为 $\Psi_i(j)=\arg\max \delta_{t-1}(i)\cdot a_{ij} \quad t\in\{1...N\}$,

依惠特比算法经递归得最终状态为 $q_N^*=\arg\max_{q_N\in S}\delta_N(q_N)$,

回溯最优路径，确定最可能的胚胎预定区域的碱基信息为 $q_t^*=\Psi_{t+1}(q^*_{t+1})$ $(t=1,2,3,...,N-1)$。

其中，这里所使用的术语“局部概率$\delta_i(q_i)$”和“逆向指针$\Psi_i(q_i)$”都是沿用 Viterbi 算法的经典定义。关于该参数的定义的详细描述，可以参见 Lawrence R. Rabiner, PROCEEDINGS OF THE IEEE， Vol.77，No.2，1989 年 2 月，通过参照将其全文并入本文。

由此，能够有效地对胚胎基因组的序列进行分析。相比其他已有的植入前检测技术方法，本方法有以下技术优势，主要体现在准确性和可获得的遗传信息量上:

1) 根据本发明实施例，针对胚胎的父源性位点和母源性位点，均可以很好地检测，检测准确率可高达 95%以上，且可以检测多种变异类型，扩大了疾病检测的范围。

2) 根据本发明实施例，对于一些在胚胎单细胞或微量细胞全基因组扩增 DNA 测序结果中覆盖程度较低，存在杂合缺失或等位基因脱扣(Allele Drop Out)的位点或 DNA 序列，可以通过本方法准确地推断得到。因此通过本方法，可以对杂合缺失或等位基因脱扣(Allele Drop Out)的位点或序列进行校正，使检测结果更加准确可靠。

3) 根据本发明实施例，可进行遗传疾病作图，对于一些连锁相关疾病，可通过其他位点的信息直接推断出来，一次可获得的信息量大，对临床检测更加具有指导意义。

另外，根据本发明实施例，本发明的确定胚胎基因组中预定区域碱基信息的方法，不仅限于 SNP 或者 STR 等某一种遗传多态性位点，对所有的遗传多态性位点均可适用，且可以多种位点同时使用，以便互相验证。

用于确定胚胎基因组中预定区域碱基信息的系统

在本发明的又一方面，本发明提出了一种用于确定胚胎基因组中预定区域核酸序列的系统。根据本发明的实施例，参考图 2，该系统 1000 可以包括: 文库构建装置 100、测序装置 200 以及分析装置 400。

根据本发明的实施例，文库构建装置 100 适于针对胚胎基因组 DNA 样本以及胚胎遗传相关个体的基因组 DNA 样本，分别构建测序文库。根据本发明的实施例，测序装置 200 与文库构建装置 100 相连，并且适于对所述测序文库进行测序，以便获得胚胎以及胚胎遗传相关个体的测序结果。根据本发明的实施例，还可以进一步包括 DNA 样本分离装置和 DNA 扩增装置（图中未示出），该 DNA 样本分离装置适于针对胚胎进行活检取样，获得胚胎细胞。胚胎细胞可来自于卵裂球、囊胚滋养层、受精卵和配子中的任一种，可以是单个细胞，也可以是微量细胞例如 2-10 个细胞，也可以采用任何含有胚胎核酸的孕妇样本。该 DNA 扩增装置适于针对活检取样获得的胚胎细胞进行全基因组扩增，以获得足够量 DNA 用于后续检测分析。由此，可以在不影响胚胎正常发育的前提下，对胚胎的基因组进行有效监测。关于全基因组扩增的方法和流程，主要包括多重链置换扩增(Multiple Displacement Amplification, MDA)和基于 PCR 的 WGA 技术，可以采用自主开发的 WGA 扩增流程，也可以采用商业化的试剂盒如 Qiagen 公司的 REPLI-g 系列的试剂盒、Sigma-Aldrich 公司的 GenomePlex WGA 试剂盒(WGA4)、Rubicon Genomics 公司的 PicoPlex WGA 试剂盒、GE-Healthcare 公司的 illustra Genomiphi WGA 试剂盒等等。

关于针对核酸样本，构建测序文库的方法和流程，本领域技术人员可以根据不同的测序技术进行适当选择，关于流程的细节，可以参见测序仪器的厂商例如 Illumina 公司所提供的规程，例如参见 Illumina 公司 Multiplexing Sample Preparation Guide （Part#1005361; Feb 2010）或 Paired-End SamplePrep Guide（Part#1005063; Feb 2010），通过参照将其并入本文。根据本发明的实施例，从生物样本提取核酸样本的方法和设备，也不受特别限制，可以采用商品化的核酸提取试剂盒进行。根据本发明的实施例，可以用于进行测序的方法和设备并不受特别限制，包括但不限于双脱氧链终止法；优选高通量的测序方法，由此，能够利用这些测序装置的高通量、深度测序的特点，进一步提高测序效率。从而，能够提高后续对测序数据进行分析，尤其是统计检验分析时的精确性和准确度。所述高通量的测序方法包括但不限于第二代测序技术或者是单分子测序技术。所述第二代测序平台（Metzker ML. Sequencing technologies-the next generation. Nat Rev Genet. 2010 Jan;11(1):31-46）包括但不限于 Illumina-Solexa（GA™、HiSeq™、Miseq 等）、ABI-SOLiD、Life Technologies-Ion Torrent/Proton 和 Roche-454（焦磷酸测序）测序平台；单分子测序平台（技术）包括但不限于 Helicos 公司的真实单分子测序技术（True Single Molecule DNA sequencing），Pacific Biosciences 公司单分子实时测序（single molecule real-time (SMRT™)），以及 Oxford Nanopore Technologies 公司的纳米孔测序技术等（Rusk, Nicole (2009-04-01). Cheap Third-Generation Sequencing. Nature Methods 6 (4): 244-245）。随着测序技术的不断进化，本领域技术人员能够理解的是还可以采用其他的测序方法和装置进行全基因组测序。根据本发明的具体示例，可以利用选自 Illumina-Solexa、ABI-SOLiD、Life Technologies-Ion Torrent/Proton、Roche-454 和单分子测序装置的至少一种对所述全基因组测序文库进行测序。

根据本发明的实施例，还可以包括比对装置 300。根据本发明的实施例，比对装置 300 与测序装置 200 相连，并且适于将所得到的测序结果与参考序列进行比对，以便通过将胚胎细胞基因组 DNA 样本的测序结果与参考序列进行比对，构建所述胚胎的遗传草图；通过将所述胚胎遗传相关个体基因组样本的测序结果与参考序列进行比对，确定所述胚胎遗传相关个体的基因型；以及基于所述胚胎遗传相关个体的基因型，确定所述胚胎父母的单倍体型。根据本发明的实施例，可以使用的参考序列的类型并不受特别限制，可以为任何含有感兴趣区域的已知序列。根据本发明的实施例，可以采用已知的人类参考基因组作为参考序列。例如，根据本发明的实施例，采用的人类参考基因组为 NCBI 36.3，HG18。另外，根据本发明的实施例，进行比对的方法并不受特别限制。根据本发明的具体实施例，可以采用 SOAP 进行比对。

根据本发明的实施例，分析装置 400 适于：基于胚胎细胞基因组 DNA 样本的测序结果，构建所述胚胎的遗传草图，以便确定胚胎初始基因型；基于所述胚胎遗传相关个体基因组样本的测序结果，确定胚胎父母的单倍体型；以及根据隐马尔可夫模型，以所述胚胎初始基因型作为观察序列，基于所述胚胎父母的单倍体型，确定胚胎基因组中预定区域的碱基信息。

根据本发明的实施例，根据隐马尔可夫模型，确定胚胎基因组中预定区域的碱基信息

进一步包括:

构建初始状态概率分布向量、隐藏状态转移的概率矩阵以及观察序列概率矩阵;

利用惠特比算法确定最终状态并回溯最优路径，以便确定胚胎基因组中预定区域的碱基信息。

根据具体的实施例，所述隐马尔可夫模型采用下列参数:

初始状态概率分布为 $\pi=\{\pi_i\},(i\in S,\pi_i=0.5)$,

隐藏状态转移的概率矩阵为 $A=\{a_{ij}\},(i,j\in S)$，其中，$a_{ij}=\begin{cases}1-Nr/Np & i=j\\ Nr/Np & i\neq j\end{cases}$

Nr, Np 分别表示期望重组数和单核苷酸多态性位点数，Nr 可取 20-40 之间的自然数，

观察序列概率矩阵为 $B=\{b_i(k)\},(i\in S,k\in V)$，其中，

$b_i(0)=\frac{\#sites(L>0,\text{Must-hom.})}{\#sites(L>0,\text{Must-het. or Must-hom.})}$, $b_i(1)=1-b_i(0)$,

#sites (L>0,Must-hom.)为子代一定为纯合的位点的数目，#sites (L>0,Must-hom. or Must-het.)为子代一定为纯合的位点的数目以及子代一定为杂合的位点的数目的总和;

局部概率为 $\delta_t(j)=\max\left[\delta_{t-1}(i)\cdot a_{ij}\right]\cdot b_j(K)\quad t\in\{1...N\}$,

逆向指针为 $\Psi_i(j)=\arg\max\delta_{t-1}(i)\cdot a_{ij}\quad t\in\{1...N\}$,

递归得最终状态为 $q_N^*=\arg\max_{q_N\in S}\delta_N(q_N)$,

回溯最优路径，确定最可能的胚胎预定区域的碱基信息为 $q_t^*=\Psi_{t+1}(q^*_{t+1})$ $(t=1,2,3,...,N-1)$。

这里所使用的术语局部概率 $\delta_i(q_i)$和逆向指针 $\Psi_i(q_i)$都是沿用 Viterbi 算法的经典定义。关于该参数的定义的详细描述，可以参见 Lawrence R. Rabiner, PROCEEDINGS OF THE IEEE， Vol.77，No.2，1989 年 2 月，通过参照将其全文并入本文。

关于数据分析部分，前面已经进行了详细描述，也当然地适用于确定胚胎基因组中预定区域核酸序列的系统。不再赘述。

由此，利用该系统，能够有效地实施前面所述的确定胚胎基因组中预定区域核酸序列的方法，可以借助隐马尔可夫模型，通过例如惠特比算法（Viterbi algorithm）解码，确定胚胎基因组中特定区域的碱基信息，由此，可以有效地对胚胎基因组的遗传信息进行植入前检测。

此外，根据本发明的实施例，预定区域为已知存在遗传多态性的位点，而遗传多态性为选自单核苷酸多态性和 STR 的至少一种。

在本文中所述的术语“相连”应作广义理解，既可以是直接相连，也可以是间接相连，只要能够实现上述功能上的衔接即可。

需要说明的是，本领域技术人员能够理解，在前面所描述的确定胚胎基因组中预定区域核酸序列的方法的特征和优点也适合于确定胚胎基因组中预定区域核酸序列的系统，为描述方便，不再详述。

计算机可读介质

在本发明的又一方面，本发明提出了一种计算机可读介质。根据本发明的实施例，所述计算机可读介质上存储有指令，所述指令适于被处理器执行以便:

基于胚胎细胞基因组 DNA 样本的测序结果，构建胚胎的遗传草图，以便确定胚胎初始基因型;

基于该胚胎遗传相关个体基因组样本的测序结果，确定胚胎父母的单倍体型；以及

根据隐马尔可夫模型，以该胚胎初始基因型作为观察序列，基于该胚胎父母的单倍体型，确定胚胎基因组中预定区域的碱基信息。

由此，利用该计算机可读介质，能够有效地实施前面所述的方法，从而可以通过例如惠特比算法（Viterbi algorithm），借助隐马尔可夫模型，确定胚胎基因组中特定区域的碱基信息，由此，可以有效地对胚胎基因组的遗传信息进行植入前检测。

根据本发明的实施例，根据隐马尔可夫模型，确定胚胎基因组中预定区域的碱基信息进一步包括:

构建初始状态概率分布向量、隐藏状态转移的概率矩阵以及观察序列概率矩阵;

利用惠特比算法确定最终状态并回溯最优路径，以便确定胚胎基因组中预定区域的碱基信息。

根据具体的实施例，所述隐马尔可夫模型采用下列参数:

初始状态概率分布为 $\pi=\{\pi_i\},(i\in S,\pi_i=0.5)$，

隐藏状态转移的概率矩阵为 $A=\{a_{ij}\},(i,j\in S)$，其中，$a_{ij}=\begin{cases}1-Nr/Np & i=j\\ Nr/Np & i\neq j\end{cases}$

Nr、Np 分别表示期望重组数和单核苷酸多态性位点数，Nr 可取 20-40 之间的自然数，观察序列概率矩阵为 $B=\{b_i(k)\},(i\in S,k\in V)$，其中，$b_i(0)=\frac{\#sites(L>0,\text{Must-hom.})}{\#sites(L>0,\text{Must-het. or Must-hom.})}$，$b_i(1)=1-b_i(0)$，

#sites (L>0,Must-hom.)为子代一定为纯合的位点的数目，#sites (L>0,Must-hom. or Must-het.)为子代一定为纯合的位点的数目以及子代一定为杂合的位点的数目的总和;

局部概率为 $\delta_t(j)=\max\left[\delta_{t-1}(i)\cdot a_{ij}\right]\cdot b_j(K) \quad t\in\{1...N\}$

逆向指针为 $\Psi_i(j)=\arg\max\delta_{t-1}(i)\cdot a_{ij} \quad t\in\{1...N\}$，

经递归得最终状态为 $q_N^*=\arg\max_{q_N\in S}\delta_N(q_N)$，

回溯最优路径，确定最可能的胚胎预定区域的碱基信息为 $q_t^*=\Psi_{t+1}(q^*_{t+1})$ $(t=1,2,3,...,N-1)$。

这里所使用的术语局部概率$\delta_i(q_i)$和逆向指针$\Psi_i(q_i)$都是沿用 Viterbi 算法的经典定义。关于该参数的定义的详细描述，可以参见 Lawrence R. Rabiner, PROCEEDINGS OF THE IEEE， Vol.77，No.2，1989 年 2 月，通过参照将其全文并入本文。

关于数据分析部分，前面已经进行了详细描述，也当然地适用于确定胚胎基因组中预定区域核酸序列的系统。不再赘述。

由此，利用该系统，能够有效地实施前面所述的确定胚胎基因组中预定区域核酸序列的方法，可以通过例如惠特比算法（Viterbi algorithm），借助隐马尔可夫模型，确定胚胎基因组中特定区域的碱基信息，由此，可以有效地对胚胎基因组的遗传信息进行植入前检测。

此外，根据本发明的实施例，预定区域为已知存在遗传多态性的位点，而遗传多态性为选自单核苷酸多态性和 STR 的至少一种。

就本说明书而言，“计算机可读介质”可以是任何可以包含、存储、通信、传播或传输程序以供指令执行系统、装置或设备或结合这些指令执行系统、装置或设备而使用的装置。计算机可读介质的更具体的示例（非穷尽性列表）包括以下：具有一个或多个布线的电连接部（电子装置），便携式计算机盘盒（磁装置），随机存取存储器（RAM），只读存储器（ROM），可擦除可编辑只读存储器（EPROM 或闪速存储器），光纤装置，以及便携式光盘只读存储器（CDROM）。另外，计算机可读介质甚至可以是可在其上打印所述程序的纸或其他合适的介质，因为例如可以通过对纸或其他介质进行光学扫描，接着进行编辑、解译或必要时以其他合适方式进行处理来以电子方式获得所述程序，然后将其存储在计算机存储器中。

应当理解，本发明的各部分可以用硬件、软件、固件或它们的组合来实现。在上述实施方式中，多个步骤或方法可以用存储在存储器中且由合适的指令执行系统执行的软件或固件来实现。例如，如果用硬件来实现，和在另一实施方式中一样，可用本领域公知的下列技术中的任一项或他们的组合来实现：具有用于对数据信号实现逻辑功能的逻辑门电路的离散逻辑电路，具有合适的组合逻辑门电路的专用集成电路，可编程门阵列（PGA），现场可编程门阵列（FPGA）等。

本技术领域的普通技术人员可以理解实现上述实施例方法携带的全部或部分步骤是可以通过程序来指令相关的硬件完成，所述的程序可以存储于一种计算机可读存储介质中，

该程序在执行时，包括方法实施例的步骤之一或其组合。

此外，在本发明各个实施例中的各功能单元可以集成在一个处理模块中，也可以是各个单元单独物理存在，也可以两个或两个以上单元集成在一个模块中。上述集成的模块既可以采用硬件的形式实现，也可以采用软件功能模块的形式实现。所述集成的模块如果以软件功能模块的形式实现并作为独立的产品销售或使用时，也可以存储在一个计算机可读取存储介质中。

下面将结合实施例对本发明的方案进行解释。本领域技术人员将会理解，下面的实施例仅用于说明本发明，而不应视为限定本发明的范围。实施例中未注明具体技术或条件的，按照本领域内的文献所描述的技术或条件（例如参考 J.萨姆布鲁克等著，黄培堂等译的《分子克隆实验指南》，第三版，科学出版社）或者按照产品说明书进行。所用试剂或仪器未注明生产厂商者，均为可以通过市购获得的常规产品，例如可以采购自 Illumina 公司。

一般方法:

参考图 1，本发明实施例的主要步骤包括:

1) 获得单细胞或微量细胞样本，第三天卵裂球细胞或第五天囊胚滋养层细胞活检取样，或受精卵取样，或其他单细胞或微量细胞取样。
2) 单细胞或微量细胞样本进行全基因组扩增，扩增产物 DNA 根据新一代测序平台要求构建文库，并进行测序。
3) 单细胞或微量细胞测序数据经过过滤后与人类基因组参考序列进行比对。
4) 根据比对结果进行胚胎细胞单核苷酸多态性和缺失重复等遗传变异的初步检测。
5) 构建胚胎基因组的遗传草图，对胚胎基因型进行初始化。
6) 收集胚胎遗传相关个体，如家系成员如父母和先证者，和/或祖父母、外祖父母等或者同一对父母的其它胚胎样本，提取基因组 DNA，根据新一代测序平台要求构建文库，并进行测序。
7) 胚胎遗传相关个体 DNA 测序数据进行过滤，并与人类基因组参考序列比对。
8) 根据比对结果，分析确定先证者、父母等遗传相关个体的基因型。
9) 利用先证者、父母等遗传相关个体的基因型，推断父母双亲的单倍体型。
10) 借助隐马尔科夫模型解码方式，用父母单倍体型结果推断胚胎单体型。
11) 胚胎遗传变异的最终确定。

参考图 1，本发明实施例中信息分析部分用到的隐马尔科夫模型以及详细分析步骤如下:

标记:

1、需要检测的位点数为 N。

2、观测序列:

3、隐藏状态集为: $S=\{0,1\}$，定义了父母哪条染色体遗传给了子代。0 代表遗传给了先

证者的那条染色体。

4、观测状态集为: $V=\{0,1\}$。1 代表遗传给先证者的染色体与胚胎遗传草图基因型一致。0 代表不一致。

第一步，构建初始状态概率分布向量，单倍体重组转移矩阵以及观察序列概率矩阵

I. 初始状态概率分布记为 $\pi=\{\pi_i\},(i\in S,\pi_i=0.5)$

在没有参考数据的情况下，设每种隐藏状态出现的可能性相等。

II. 记单倍体重组转移矩阵记为 $A=\{a_{ij}\},(i,j\in S)$,即隐藏状态转移的概率。

$$a_{ij}=\begin{cases}1-Nr/Np & i=j\\ Nr/Np & i\neq j\end{cases}$$

Nr, Np 分别表示期望重组数和单核苷酸多态性位点数。

构建观察序列概率矩阵

记观察序列概率矩阵为 $B=\{b_i(k)\},(i\in S,k\in V)$，这里的概率分布用子代一定是纯合（Must-hom）或一定是杂合(Must-het)的位点来估算。

$$b_i(0)=\frac{\#sites(L>0,\text{Must-hom.})}{\#sites(L>0,\text{Must-het. or Must-hom.})},\ b_i(1)=1-b_i(0)$$

第二步，构建局部概率矩阵，和逆向指针

定义 局部概率 : $\delta_t(j)=\max\left[\delta_{t-1}(i)\cdot a_{ij}\right]\cdot b_j(K)\quad t\in\{1...N\}$

定义 逆向指针 : $\Psi_t(j)=\arg\max\delta_{t-1}(i)\cdot a_{ij}\quad t\in\{1...N\}$

第三步，确定最终状态状态，并回溯最优路径

确定最终状态，$q_N^*=\arg\max_{q_N\in S}\delta_N(q_N)$

回溯最优路径，即最可能胚胎单基因型 $q_t^*=\Psi_{t+1}(q^*_{t+1})$ $(t=1,2,3,...,N-1)$

第四步，输出结果

实施例 1

一、样品收集及处理:

一对夫妇曾生育过一个重度β地中海贫血的女儿 A(先证者)，因此接受 IVF（体外受精联合胚胎移植技术）-PGD 治疗，希望能再生育一个与先证者 HLA 配型相符的健康宝宝。经过一个 IVF-PGD 治疗周期，在第三天进行卵裂球胚胎单细胞活检取样，活检后的卵裂球继续培养至第五天囊胚期，进行囊胚玻璃化冷冻保存。活检获得的卵裂球单细胞样本进行

全基因组扩增(PED7-1-OWGA)，全基因组扩增采用 Qiagen 公司的 REPLI-g MDA 扩增试剂盒，严格根据试剂盒说明书进行操作，扩增产物用于新一代测序及其他分析。同时也收集这个家系中父亲(PED4-F)、母亲(PED3-M)和先证者女儿(PED-Sister)的外周血，提取基因组 DNA。

在完成 IVF-PGD 治疗后，选择正确的胚胎植入子宫，成功妊娠后，可以在 14 周以后抽取孕妇羊水(PED7-2-OAF)，或收集胚胎脐带血，提取 DNA，进行测序，以验证胚胎细胞的分型结果。

胚胎单细胞全基因组扩增产物 DNA 用 Covaris™ 打断仪打断至 200bp 大小的片段，根据 illumia®公司 HiSeq2000™ 测序仪的上机要求进行建库，构建好的文库经 Agilent®Bioanalyzer 2100 和 Q-PCR 方法进行质控，合格的文库继续进行 illumina® HiSeq2000™ 测序仪测序，测序策略为 Pair End 90 index（即双向 90bp index 测序），测序深度 30X，全基因组测序。其中仪器的参数设置及操作方法都按照 illumina®操作手册（可由 http://www.illumina.com/support/documentation.ilmn 获取）。

父亲、母亲和先证者女儿外周血 DNA，以及孕妇羊水或胚胎脐带血 DNA 用 Covaris™ 打断仪打断至 200bp 大小的片段，根据 illumina®公司 HiSeq2000™ 测序仪的上机要求进行建库，文库进行 All in One 芯片捕获，具体操作见 NimbleGen SeqCap EZ 芯片捕获建库说明书(http://www.nimblegen.com/products/seqcap/ez/index.html 可获取)。All in One 芯片是自主设计的目标区域捕获芯片，其目标区域包括全外显子区域、1M 的 SNP 位点区域和 MHC 基因区。芯片捕获后的文库经 Agilent®Bioanalyzer 2100 和 Q-PCR 方法进行质控，合格的文库继续进行 illumina® HiSeq2000™ 测序仪测序，测序策略为 Pair End 90 index，测序深度为目标区域的 30-50X。其中仪器的参数设置及操作方法都按照 illumina®操作手册（可由 http://www.illumina.com/support/documentation.ilmn 获取）。

二、测序数据分析:

1、先证者和父母基因组 DNA 测序分型:

a) 过滤掉低质量值和接头污染的序列

b) 使用 Burrows-Wheeler Aligner (BWA, http://bio-bwa.sourceforge.net/)将高质量的测序序列比对在人类参考基因组（Hg19, NCBI release GRCh37)。并用 SAMtools 去除掉 PCR 重复扩增的序列。

c) 使用 Genome Analysis Toolkit (GATK, http://www.broadinstitute.org/gatk/)校准碱基质量值并进行单碱基核苷酸多态性和缺失重复的检测。

2、胚胎基因组遗传草图构建。

3、根据先证者基因型推断父母的单倍体型:

根据核心家系推断单体型，具体可参考文献 B. L. Browning and S. R. Browning. Am J Hum Genet 84:210-223.(2009)进行。

4、根据父母单体型推断胚胎细胞的单倍体型：

a) 使用 BWA 将胚胎单细胞测序数据比对到人类参考基因组（Hg19, NCBI release GRCh37)

b) 构建初始状态概率向量，以及胚胎单倍体重组转移矩阵，Nr=30，其它按上所述。

c) 统计胚胎细胞每个位点的测序信息，并构建观察序列概率矩阵（按上所述）。

d) 构建局部概率矩阵，和逆向指针（按上所述）。

e) 确定最终状态，并回溯最优路径

f) 输出

g) 以胚胎出生后的羊水基因分型结果为参照，对胚胎细胞的基因分型准确性统计如下：

种类		常染色体			性染色体		
		准确数	位点数	准确率	准确数	位点数	准确率
重构前	纯合	175,194	233,794	74.94	2,345	3,318	70.68
	杂合	81,129	81,576	99.45	1,476	1,480	99.72
重构后	纯合	172,968	173,631	99.62	2,329	2,330	99.96
	草图杂合	80,877	81,042	99.80	1,472	1,478	99.59
	更正杂合	58,194	60,194	95.88	967	990	97.68

注："草图杂合"表示重构前后均为杂合的位点；"更正杂合"表示重构前为纯合，重构后为杂合的位点。

5、β-地贫和 HLA 型别判断

a) 通过父母单体型，成功还原了胚胎细胞在 rs7480526 上的基因型，预测其为β-地贫阴性，并发现与先证者在该位置上的等位基因不一样，从而确定胚胎为β-地贫阴性。

b) 通过本方法可以得出，先证者和胚胎在 MHC 区的单体型一样，从而判断她们的 HLA 型别是匹配的。

根据上述胚胎的β-地贫基因型和 HLA 分型结果，从而选择正确的胚胎，即β-地贫阴性和 HLA 型别与先证者匹配的胚胎，进行冷冻复苏和植入子宫，从而防止患病儿的妊娠。

实施例 2

一、样品收集及处理：

一对夫妇均为β-地贫携带者，接受 IVF-PGD 治疗。经过一个 IVF-PGD 治疗周期，在第三天获得了 8 个卵裂球胚胎，均进行卵裂球胚胎单细胞活检取样，活检后的卵裂球继续培养至第五天囊胚期，进行囊胚玻璃化冷冻保存。活检获得的 8 个卵裂球单细胞样本进行全基因组扩增，全基因组扩增采用 Qiagen 公司的 REPLI-g MDA 扩增试剂盒，严格根据试剂盒说明书进行操作，扩增产物用于新一代测序及其他分析。

同时也收集这个家系中父亲、母亲的外周血，提取基因组 DNA。

在完成 IVF-PGD 治疗，选择正确的胚胎植入子宫，成功妊娠后，可以在 14 周以后抽取孕妇羊水，或收集胚胎脐带血，提取 DNA，进行测序，以验证胚胎细胞的分型结果。

8 个胚胎单细胞全基因组扩增产物 DNA，父亲、母亲外周血 DNA，以及孕妇羊水或胚胎脐带血 DNA 用 Covaris™ 打断仪打断至 200bp 大小的片段，根据 illumina®公司 HiSeq2000™ 测序仪的上机要求进行建库，文库进行 All in One 芯片捕获，具体操作见 NimbleGen SeqCap EZ 芯片捕获建库说明书 (http://www.nimblegen.com/products/seqcap/ez/index.html 可获取)。All in One 芯片是自主设计的目标区域捕获芯片，其目标区域包括全外显子区域、1M 的 SNP 位点区域和 MHC 基因区。芯片捕获后的文库经 Agilent®Bioanalyzer 2100 和 Q-PCR 方法进行质控，合格的文库继续进行 illumina® HiSeq2000™ 测序仪测序，测序策略为 Pair End 90 index（即双向 90bp index 测序），测序深度为目标区域的 30-50X。其中仪器的参数设置及操作方法都按照 illumina® 操作手册（可由 http://www.illumina.com/support/documentation.ilmn 获取）。

二、测序数据分析:

1) 父母基因组 DNA，羊水 DNA（验证样本）测序数据分析和分型:

a) 过滤掉低质量值和接头污染的序列

b) 使用 Burrows-Wheeler Aligner (BWA, http://bio-bwa.sourceforge.net/)将高质量的测序序列比对在人类参考基因组（Hg19, NCBI release GRCh37)。并用 SAMtools 去除掉 PCR 重复扩增的序列。

c) 使用 Genome Analysis Toolkit (GATK, http://www.broadinstitute.org/gatk/)校准碱基质量值并进行单碱基核苷酸多态性和缺失重复的检测。

2) 根据 8 个胚胎单细胞测序数据分析，分别给出每个胚胎细胞基因组初始化的基因型信息，构建每个胚胎的细胞基因组遗传草图。

3）确定 8 个胚胎单细胞任一个待测胚胎，根据待测胚胎细胞基因组遗传草图和父母的基因型，对父母的单倍体型进行初始化。

4）借助隐马尔科夫模型，分别根据其余 7 个胚胎的基因型重构并确定父母的单倍体型。

5）借助隐马尔科夫模型，根据父母单倍体型，来重构待测胚胎的单体型。

6）待测胚胎遗传变异信息的确定。

以 22 号染色体为例，参照羊水基因分型结果，对待测胚胎的基因分型准确性统计如下:

类型(母 x 父)	总位点数	SNP Call 率	准确率(杂合)	准确率(纯合)
纯合 x 杂合	2183	97.16%	96.68%	99.63%
杂合 x 纯合	2156	93.18%	98.95%	100%
杂合 x 杂合	1154	82.32%	99.31%	97.59%

7) β-地贫和 HLA 型别判断

a) 通过父母单体型，成功还原了待测胚胎细胞在 rs7480526 上的单体型，预测其为 β-地贫阴性。

b) 通过本方法，结合先症者的遗传检测结果，可判定先证者和胚胎在 MHC 区的单体型是否一致，从而判断她们的 HLA 型别是否匹配。

根据上述胚胎的β-地贫基因型和 HLA 分型结果，从而选择正确的胚胎，即β-地贫阴性和 HLA 型别与先症者匹配的胚胎，进行冷冻复苏和植入子宫，从而防止患病儿的妊娠。

工业实用性

本发明的确定胚胎基因组中预定区域碱基信息的方法、用于确定胚胎基因组中预定区域碱基信息的系统以及计算机可读介质，能够有效地应用于对胚胎基因组中预定区域的核酸序列进行分析。

尽管本发明的具体实施方式已经得到详细的描述，本领域技术人员将会理解。根据已经公开的所有教导，可以对那些细节进行各种修改和替换，这些改变均在本发明的保护范围之内。本发明的全部范围由所附权利要求及其任何等同物给出。

在本说明书的描述中，参考术语“一个实施例”、“一些实施例”、“示意性实施例”、“示例”、“具体示例”、或“一些示例”等的描述意指结合该实施例或示例描述的具体特征、结构、材料或者特点包含于本发明的至少一个实施例或示例中。在本说明书中，对上述术语的示意性表述不一定指的是相同的实施例或示例。而且，描述的具体特征、结构、材料或者特点可以在任何的一个或多个实施例或示例中以合适的方式结合。

权利要求书

1、一种确定胚胎基因组中预定区域碱基信息的方法，其特征在于，包括下列步骤：

获得胚胎细胞基因组 DNA 样本的测序结果，以及胚胎遗传相关个体基因组样本的测序结果；

基于胚胎细胞基因组 DNA 样本的测序结果，构建所述胚胎的遗传草图，以便确定胚胎初始基因型；

基于所述胚胎遗传相关个体基因组样本的测序结果，确定胚胎父母的单倍体型；以及

根据隐马尔可夫模型，以所述胚胎初始基因型作为观察序列，基于所述胚胎父母的单倍体型，确定胚胎基因组中预定区域的碱基信息。

2、根据权利要求 1 所述的方法，其特征在于，所述胚胎基因组 DNA 样本是从微量胚胎细胞获得的，其中，所述微量胚胎细胞来自于卵裂球、囊胚或受精卵。

3、根据权利要求 2 所述的方法，其特征在于，所述胚胎基因组 DNA 样本是从胚胎单细胞获得的。

4、根据权利要求 1 所述的方法，其特征在于，所述测序结果是利用选自 Illumina-Solexa/Hiseq、ABI-SOLiD、Life Technologies-Ion Torrent/Proton、Roche-454 和单分子测序装置的至少一种获得的。

5、根据权利要求 1 所述的方法，其特征在于，通过将胚胎细胞基因组 DNA 样本的测序结果与参考序列进行比对，构建所述胚胎的遗传草图。

6、根据权利要求 1 所述的方法，其特征在于，通过将所述胚胎遗传相关个体基因组样本的测序结果与参考序列进行比对，确定所述胚胎遗传相关个体的基因型；以及

基于所述胚胎遗传相关个体的基因型，确定所述胚胎父母的单倍体型。

7、根据权利要求 5 或 6 所述的方法，其特征在于，所述参考序列为人类参考基因组。

8、根据权利要求 1 所述的方法，其特征在于，所述胚胎遗传相关个体包括所述胚胎的父母以及先证者。

9、根据权利要求 8 所述的方法，其特征在于，所述先证者为所述胚胎父母的其他子女。

10、根据权利要求 9 所述的方法，其特征在于，根据隐马尔可夫模型，确定胚胎基因组中预定区域的碱基信息进一步包括：

构建初始状态概率分布向量、隐藏状态转移的概率矩阵以及观察序列概率矩阵；

利用惠特比算法确定最终状态并回溯最优路径，以便确定胚胎基因组中预定区域的碱基信息。

11、根据权利要求 10 所述的方法，其特征在于，所述隐马尔可夫模型采用下列参数：

初始状态概率分布为 $\pi = \{\pi_i\}, (i \in S, \pi_i = 0.5)$，

隐藏状态转移的概率矩阵为 $A=\{a_{ij}\},(i,j\in S)$，其中，$a_{ij}=\begin{cases}1-Nr/Np & i=j\\ Nr/Np & i\neq j\end{cases}$

Nr、Np 分别表示期望重组数和单核苷酸多态性位点数，Nr 为自然数，取值范围 20-40，观察序列概率矩阵为 $B=\{b_i(k)\},(i\in S,k\in V)$，其中，$b_i(0)=\frac{\#sites(L>0,\text{Must-hom.})}{\#sites(L>0,\text{Must-het. or Must-hom.})}$，$b_i(1)=1-b_i(0)$，

#sites (L>0,Must-hom.)为子代一定为纯合的位点的数目，#sites (L>0,Must-hom. or Must-het.)为子代一定为纯合的位点的数目以及子代一定为杂合的位点的数目的总和；

局部概率为 $\delta_t(j)=\max\left[\delta_{t-1}(i)\cdot a_{ij}\right]\cdot b_j(K)\quad t\in\{1...N\}$，

逆向指针为 $\Psi_t(j)=\arg\max\delta_{t-1}(i)\cdot a_{ij}\quad t\in\{1...N\}$，

递归得最终状态为 $q_N^*=\arg\max_{q_N\in S}\delta_N(q_N)$，

回溯最优路径，确定最可能的胚胎预定区域的碱基信息为 $q_t^*=\Psi_{t+1}(q^*_{t+1})$ $(t=1,2,3,...,N-1)$。

12、根据权利要求 1 所述的方法，其特征在于，进一步包括：所述预定区域为已知存在遗传多态性的位点。

13、根据权利要求 12 所述的方法，其特征在于，所述遗传多态性为选自单核苷酸多态性和 STR 的至少一种。

14、一种用于确定胚胎基因组中预定区域碱基信息的系统，其特征在于，包括：

文库构建装置，所述文库构建装置适于针对胚胎基因组 DNA 样本以及胚胎遗传相关个体的基因组 DNA 样本，分别构建测序文库；

测序装置，所述测序装置与所述文库构建装置相连，并且适于对所述测序文库进行测序，以便获得胚胎以及胚胎遗传相关个体的测序结果；以及

分析装置，所述分析装置与所述测序装置相连，并且适于：

基于胚胎基因组 DNA 样本的测序结果，构建所述胚胎的遗传草图，以便确定胚胎初始基因型；

基于所述胚胎遗传相关个体基因组样本的测序结果，确定胚胎父母的单倍体型；以及

根据隐马尔可夫模型，以所述胚胎初始基因型作为观察序列，基于所述胚胎父母的单倍体型，确定胚胎基因组中预定区域的碱基信息。

15、根据权利要求 14 所述的系统，其特征在于，进一步包括 DNA 样本分离装置，所述 DNA 样本分离装置适于从胚胎进行活检取样，获得微量胚胎细胞，从所述微量胚胎细胞

提取胚胎基因组 DNA 样本，其中，所述微量胚胎细胞来自于卵裂球、囊胚或受精卵。

16、根据权利要求 14 所述的系统，其特征在于，所述测序装置为选自 Illumina-Solexa/Hiseq、ABI-SOLiD、Life Technologies-Ion Torrent/Proton、Roche-454 和单分子测序装置的至少一种。

17、根据权利要求 14 所述的系统，其特征在于，进一步包括比对装置，所述比对装置与所述测序装置相连，用于将所述测序结果与参考序列进行比对，以便:

通过将胚胎细胞基因组 DNA 样本的测序结果与参考序列进行比对，构建所述胚胎的遗传草图;

通过将所述胚胎遗传相关个体基因组样本的测序结果与参考序列进行比对，确定所述胚胎遗传相关个体的基因型；以及

基于所述胚胎遗传相关个体的基因型，确定所述胚胎父母的单倍体型。

18、根据权利要求 14 所述的系统，其特征在于，根据隐马尔可夫模型，确定胚胎基因组中预定区域的碱基信息进一步包括:

构建初始状态概率分布向量、隐藏状态转移的概率矩阵以及观察序列概率矩阵;

利用惠特比算法确定最终状态并回溯最优路径，以便确定胚胎基因组中预定区域的碱基信息。

19、根据权利要求 18 所述的系统，其特征在于，所述隐马尔可夫模型采用下列参数:

初始状态概率分布为 $\pi=\{\pi_i\},(i\in S,\pi_i=0.5)$,

隐藏状态转移的概率矩阵为 $A=\{a_{ij}\},(i,j\in S)$，其中，$a_{ij}=\begin{cases}1-Nr/Np & i=j\\ Nr/Np & i\neq j\end{cases}$

Nr、Np 分别表示期望重组数和单核苷酸多态性位点数，Nr 为自然数，取值范围 20-40，

观察序列概率矩阵为 $B=\{b_i(k)\},(i\in S,k\in V)$，其中，

$b_i(0)=\frac{\#sites(L>0,\text{Must-hom.})}{\#sites(L>0,\text{Must-het. or Must-hom.})}$, $b_i(1)=1-b_i(0)$,

#sites (L>0,Must-hom.)为子代一定为纯合的位点的数目，#sites (L>0,Must-hom. or Must-het.)为子代一定为纯合的位点的数目以及子代一定为杂合的位点的数目的总和;

局部概率为 $\delta_t(j)=\max\left[\delta_{t-1}(i)\cdot a_{ij}\right]\cdot b_j(K)\quad t\in\{1...N\}$,

逆向指针为 $\Psi_t(j)=\arg\max\delta_{t-1}(i)\cdot a_{ij}\quad t\in\{1...N\}$,

递归得最终状态为 $q_N^*=\arg\max_{q_N\in S}\delta_N(q_N)$,

回溯最优路径，确定最可能的胚胎预定区域的碱基信息为 $q_t^*=\Psi_{t+1}(q^*_{t+1})$

($t=1,2,3,...,N-1$)。

20、一种计算机可读介质，其特征在于，所述计算机可读介质上存储有指令，所述指令适于被处理器执行以便:

基于胚胎细胞基因组 DNA 样本的测序结果，构建胚胎的遗传草图，以便确定胚胎初始基因型;

基于所述胚胎遗传相关个体基因组样本的测序结果，确定胚胎父母的单倍体型；以及

根据隐马尔可夫模型，以所述胚胎初始基因型作为观察序列，基于所述胚胎父母的单倍体型，确定胚胎基因组中预定区域的碱基信息。

21、根据权利要求 20 所述的计算机介质，其特征在于，通过将胚胎细胞基因组 DNA 样本的测序结果与参考序列进行比对，构建所述胚胎的遗传草图。

22、根据权利要求 20 所述的计算机介质，其特征在于，通过将所述胚胎遗传相关个体基因组样本的测序结果与参考序列进行比对，确定所述胚胎遗传相关个体的基因型；以及

基于所述胚胎遗传相关个体的基因型，确定所述胚胎父母的单倍体型。

23、根据权利要求 20 所述的计算机介质，其特征在于，根据隐马尔可夫模型，确定胚胎基因组中预定区域的碱基信息进一步包括:

构建初始状态概率分布向量、隐藏状态转移的概率矩阵以及观察序列概率矩阵;

利用患特比算法确定最终状态并回溯最优路径，以便确定胚胎基因组中预定区域的碱基信息。

24、根据权利要求 23 所述的方法，其特征在于，所述隐马尔可夫模型采用下列参数:

初始状态概率分布为 $\pi=\{\pi_i\},(i\in S,\pi_i=0.5)$，

隐藏状态转移的概率矩阵为 $A=\{a_{ij}\},(i,j\in S)$，其中，$a_{ij}=\begin{cases}1-Nr/Np & i=j\\ Nr/Np & i\neq j\end{cases}$

Nr、Np 分别表示期望重组数和单核苷酸多态性位点数，Nr 为自然数，取值范围 20-40，

观察序列概率矩阵为 $B=\{b_i(k)\},(i\in S,k\in V)$，其中，

$b_i(0)=\dfrac{\#sites(L>0,\text{Must-hom.})}{\#sites(L>0,\text{Must-het. or Must-hom.})}$，$b_i(1)=1-b_i(0)$，

#sites (L>0,Must-hom.)为子代一定为纯合的位点的数目，#sites (L>0,Must-hom. or Must-het.)为子代一定为纯合的位点的数目以及子代一定为杂合的位点的数目的总和;

局部概率为 $\delta_t(j)=\max\left[\delta_{t-1}(i)\cdot a_{ij}\right]\cdot b_j(K)\quad t\in\{1...N\}$，

逆向指针为 $\Psi_i(j)=\arg\max\delta_{t-1}(i)\cdot a_{ij}\quad t\in\{1...N\}$，

递归得最终状态为 $q_N^*=\arg\max_{q_N\in S}\delta_N(q_N)$，

回溯最优路径，确定最可能的胚胎预定区域的碱基信息为 $q_t^* = \Psi_{t+1}(q^*_{t+1})$ ($t = 1,2,3,...,N-1$)。

25、根据权利要求 20 所述的计算机可读介质，其特征在于，所述预定区域为已知存在遗传多态性的位点。

26、根据权利要求 25 所述的计算机可读介质，其特征在于，所述遗传多态性为选自单核苷酸多态性和 STR 的至少一种。

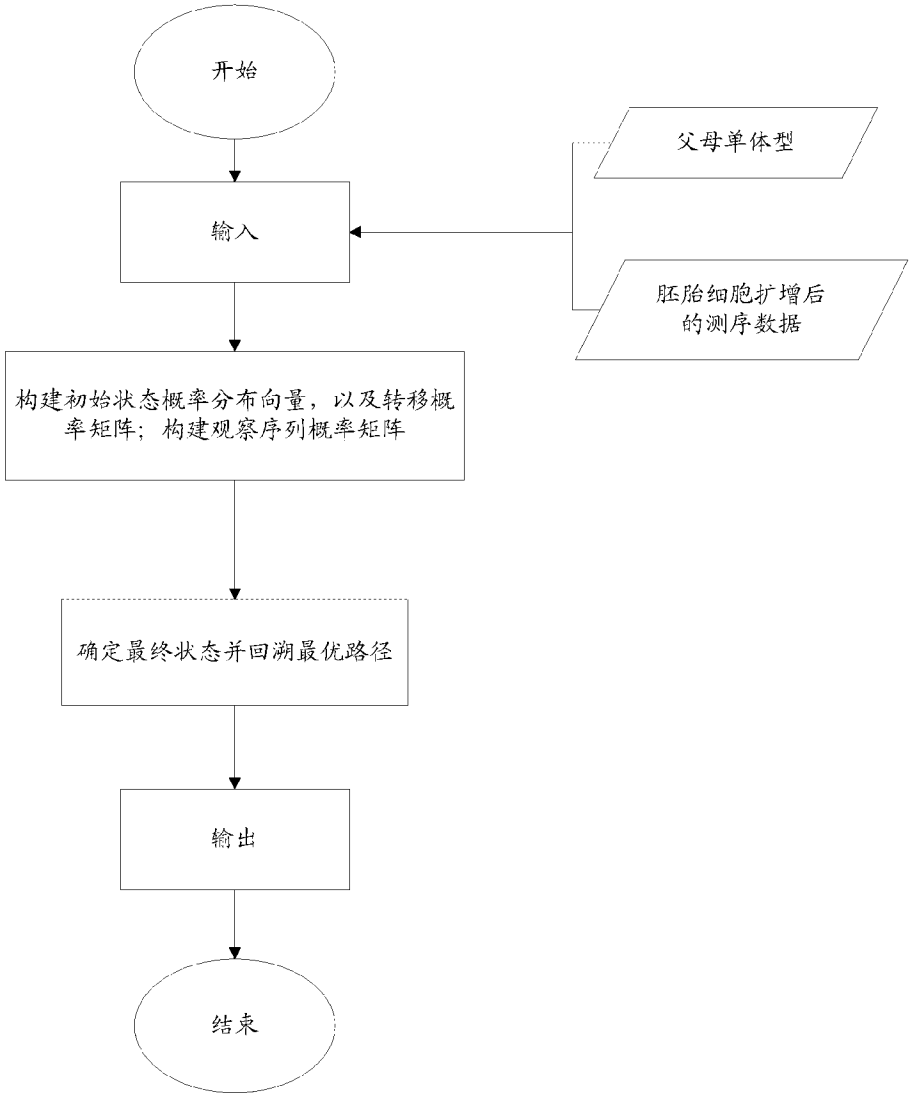

开始
输入
父母单体型
胚胎细胞扩增后的测序数据
构建初始状态概率分布向量，以及转移概率矩阵；构建观察序列概率矩阵
确定最终状态并回溯最优路径
输出
结束

图 1

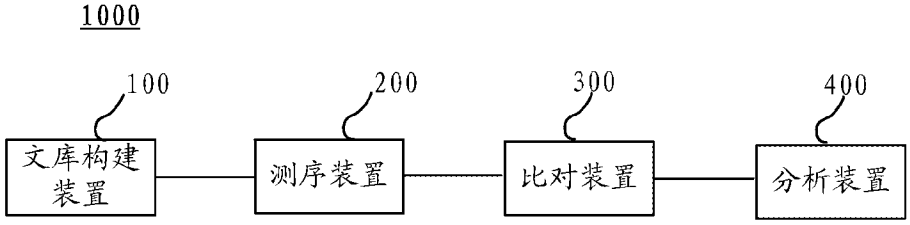

1000
100
200
300
400
文库构建装置
测序装置
比对装置
分析装置

图 2

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2013/073375

A. CLASSIFICATION OF SUBJECT MATTER

C12Q 1/68 (2006.01) i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

IPC: C12Q

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

CNPAT, WPI, EPODOC, CNKI, PubMed, GOOGLE, NCBI: hidden markov model, HMM, Viterbi algorithm, library, sequencing, hidden, embryo, creat

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	YU, Weidong et al. Construction of cDNA Library from Rabbit Single Preimplantation Cloned Embryos. Chinese Journal of Biochemistry and Molecular Biology, April 2004, vol. 20, no. 2, pages 162-165, page 163, right columns 1.2.3 and 1.2.4	1-19
Y	CN 102317473 A (PACIFIC BIOSCIENCES CALIFORNIA) 11 January 2012 (11.01.2012) see claims 1 and 2, description, paragraphs [0161]-[0166]	1-19
A	CN 102127818 A (ZHANG, Kang) 20 July 2011 (20.07.2011) see the whole document	1-19

☐ Further documents are listed in the continuation of Box C. ☒ See patent family annex.

* Special categories of cited documents:

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier application or patent but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&"document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
10 December 2013 (10.12.2013)	16 January 2014 (16.01.2014)
Name and mailing address of the ISA State Intellectual Property Office of the P. R. China No. 6, Xitucheng Road, Jimenqiao Haidian District, Beijing 100088, China Facsimile No. (86-10) 62019451	Authorized officer ZHANG, Bin Telephone No. (86-10) 61648321

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

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2013/073375

Box No. II Observations where certain claims were found unsearchable (Continuation of item 2 of first sheet)

This international search report has not been established in respect of certain claims under Article 17(2)(a) for the following reasons:

1. ☒ Claims Nos.:20-26
because they relate to subject matter not required to be searched by this Authority, namely:
The subject matter of claims 20-26 relates to mere presentations of information and therefore does not warrant an international search according to the criteria set out in PCT Rule 39. 1(iv).

2. ☐ Claims Nos.:
because they relate to parts of the international application that do not comply with the prescribed requirements to such an extent that no meaningful international search can be carried out, specifically:

3. ☐ Claims Nos.:
because they are dependent claims and are not drafted in accordance with the second and third sentences of Rule 6.4(a).

Box No. III Observations where unity of invention is lacking (Continuation of item 3 of first sheet)

This International Searching Authority found multiple inventions in this international application, as follows:

1. ☐ As all required additional search fees were timely paid by the applicant, this international search report covers all searchable claims.

2. ☐ As all searchable claims could be searched without effort justifying additional fees, this Authority did not invite payment of additional fees.

3. ☐ As only some of the required additional search fees were timely paid by the applicant, this international search report covers only those claims for which fees were paid, specifically claims Nos.:

4. ☐ No required additional search fees were timely paid by the applicant. Consequently, this international search report is restricted to the invention first mentioned in the claims; it is covered by claims Nos.:

Remark on protest

☐ The additional search fees were accompanied by the applicant's protest and, where applicable, the payment of a protest fee.

☐ The additional search fees were accompanied by the applicant's protest but the applicable protest fee was not paid within the time limit specified in the invitation.

☐ No protest accompanied the payment of additional search fees.

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2013/073375

Patent Documents referred in the Report	Publication Date	Patent Family	Publication Date
CN 102317473 A	11.01.2012	WO 2010068289 A2	17.06.2010
		US 2011183320 A1	28.07.2011
		US 2010221716 A1	02.09.2010
		CA 2746632 A1	17.06.2010
		WO 2010068289 A3	14.10.2010
		AU 2009325069 A1	17.06.2010
		EP 2370598 A2	05.10.2011
		US 2013029853 A1	31.01.2013
CN 102127818 A	20.07.2011	None	

Form PCT/ISA /210 (patent family annex) (July 2009)

国际检索报告

国际申请号
PCT/CN2013/073375

A. 主题的分类

C12Q 1/68 (2006.01)i

按照国际专利分类(IPC)或者同时按照国家分类和 IPC 两种分类

B. 检索领域

检索的最低限度文献(标明分类系统和分类号)

IPC : C12Q

包含在检索领域中的除最低限度文献以外的检索文献

在国际检索时查阅的电子数据库(数据库的名称，和使用的检索词（如使用）)

CNPAT, WPI, EPODOC, CNKI, PubMed, GOOGLE, NCBI：隐马尔可夫模型，惠特比算法，文库，测序，隐马，胚胎，构建，测序，library，sequencing，hidden，HMM.

C. 相关文件

类　型*	引用文件，必要时，指明相关段落	相关的权利要求
Y	郁卫东 等，先单个植入前克隆胚胎 cDNA 文库的构建，中国生物化学与分子生物学报，04 月 2004，第 20 卷，第 2 期，第 162-165 页 参见第 163 页右栏 1.2.3 和 1.2.4	1-19
Y	CN102317473 A（加利福尼亚太平洋生物科学股份有限公司）11.01 月 2012（11.01.2012） 参见权利要求 1-2，说明书第[0161]-[0166]段	1-19
A	CN102127818 A（张康）20.07 月 2011（20.07.2011） 参见全文	1-19

☐ 其余文件在 C 栏的续页中列出。　　☒ 见同族专利附件。

* 引用文件的具体类型：
"A" 认为不特别相关的表示了现有技术一般状态的文件
"E" 在国际申请日的当天或之后公布的在先申请或专利
"L" 可能对优先权要求构成怀疑的文件，或为确定另一篇引用文件的公布日而引用的或者因其他特殊理由而引用的文件（如具体说明的）
"O" 涉及口头公开、使用、展览或其他方式公开的文件
"P" 公布日先于国际申请日但迟于所要求的优先权日的文件
"T" 在申请日或优先权日之后公布，与申请不相抵触，但为了理解发明之理论或原理的在后文件
"X" 特别相关的文件，单独考虑该文件，认定要求保护的发明不是新颖的或不具有创造性
"Y" 特别相关的文件，当该文件与另一篇或者多篇该类文件结合并且这种结合对于本领域技术人员为显而易见时，要求保护的发明不具有创造性
"&" 同族专利的文件

国际检索实际完成的日期	国际检索报告邮寄日期
10.12 月 2013（10.12.2013）	16.1 月 2014 (16.01.2014)
ISA/CN 的名称和邮寄地址： 中华人民共和国国家知识产权局 中国北京市海淀区蓟门桥西上城路 6 号 100088 传真号：　(86-10)62019451	受权官员 张彬 电话号码：(86-10) **61648321**

PCT/ISA/210 表(第 2 页) (2009 年 7 月)

国际检索报告

国际申请号
PCT/CN2013/073375

第II栏　某些权利要求被认为是不能检索的意见(续第1页第2项)

根据条约第17条(2)(a)，对某些权利要求未做国际检索报告的理由如下：

1. ☒ 权利要求：20-26

因为它们涉及不要求本单位进行检索的主题，即：

权利要求20-26涉及单纯的信息提供（PCT细则39.1(v)）。

2. ☐ 权利要求：

因为它们涉及国际申请中不符合规定的要求的部分，以致不能进行任何有意义的国际检索，

具体地说：

3. ☐ 权利要求：

因为它们是从属权利要求，并且没有按照细则6.4(a)第2句和第3句的要求撰写。

第III栏　缺乏发明单一性的意见(续第1页第3项)

本国际检索单位在该国际申请中发现多项发明，即：

1. ☐ 由于申请人按时缴纳了被要求缴纳的全部附加检索费，本国际检索报告涉及全部可作检索的权利要求。

2. ☐ 由于无需付出有理由要求附加费的劳动即能对全部可检索的权利要求进行检索，本单位未通知缴纳任何附加费。

3. ☐ 由于申请人仅按时缴纳了部分被要求缴纳的附加检索费，本国际检索报告仅涉及已缴费的那些权利要求。

具体地说，是权利要求：

4. ☐ 申请人未按时缴纳被要求缴纳的附加检索费。因此，本国际检索报告仅涉及权利要求书中首先提及的发明；包含该发明的权利要求是：

关于异议的说明： ☐ 申请人缴纳了附加检索费，同时提交了异议书，适用时，缴纳了异议费。

☐ 申请人缴纳了附加检索费，同时提交了异议书，但未在通知书规定的时间期限内缴纳异议费。

☐ 缴纳附加检索费时未提交异议书。

国际检索报告
关于同族专利的信息

国际申请号
PCT/CN2013/073375

检索报告中引用的专利文件	公布日期	同族专利	公布日期
CN 102317473 A	11.01.2012	WO 2010068289 A2	17.06.2010
		US 2011183320 A1	28.07.2011
		US 2010221716 A1	02.09.2010
		CA 2746632 A1	17.06.2010
		WO 2010068289 A3	14.10.2010
		AU 2009325069 A1	17.06.2010
		EP 2370598 A2	05.10.2011
		US 2013029853 A1	31.01.2013
CN 102127818 A	20.07.2011	无	

Espacenet

Bibliographic data: CN104053789 (B) — 2016-02-10

Method, System And Computer Readable Medium For Determining Base Information In Predetermined Area Of Fetus Genome

Inventor(s): CHEN SHENGPEI, ; GE HUIJUAN, ; LI XUCHAO, ; YI SHANG, ; WANG JIAN, ; WANG JUN, ; YANG HUANMING, ; ZHANG XIUQIN

Applicant(s): BGI HEALTH SERVICE CO LTD

Classification: - international: *C12Q1/68*
- cooperative: **C12Q1/6809 (CN)**; **C12Q1/6858 (EP, KR, US)**; **C12Q1/6869 (EP, KR, US)**; **C12Q1/6883 (US)**; **G16B20/00 (EP, KR, US)**; **G16B30/00 (EP, KR, US)**; **G16B40/00 (KR)**; **G16H50/20 (EP, KR, US)**; **G16H50/30 (EP, KR, US)**; C12Q2600/118 (US); C12Q2600/156 (US) more

Application number: CN20128067404 20120514 Global Dossier

Priority number(s): WO2012CN75478 20120514

Also published as: CN104053789 (A) AU2012380221 (A1) AU2012380221 (B2) EP2851431 (A1) EP2851431 (A4) EP2851431 (B1) ES2656023 (T3) HK1196401 (A1) JP2015525062 (A) JP6045686 (B2) KR101770460 (B1) KR20140146193 (A) PL2851431 (T3) RU2014150655 (A) RU2597981 (C2) SG11201407515R (A) US2015094210 (A1) US2018320235 (A1) WO2013170429 (A1) less

Abstract of CN104053789 (A)

Provided are a method, system and computer readable medium for determining the base information in a predetermined area of a fetus genome, the method comprising the following steps: constructing a sequence library for the DNA samples of the fetus genome; sequencing the sequence library to obtain the sequencing result of the fetus, the sequencing result of the fetus comprised of a plurality of sequencing data; and based on the sequencing result of the fetus, determining the base information in the predetermined area according to the hidden Markov model in conjunction with the genetic information of an individual related hereditarily to the fetus.

(19)中华人民共和国国家知识产权局

(12)发明专利

(10)授权公告号 CN 104053789 B
(45)授权公告日 2016.02.10

(21)申请号 201280067404.1

(22)申请日 2012.05.14

(85)PCT国际申请进入国家阶段日
2014.07.17

(86)PCT国际申请的申请数据
PCT/CN2012/075478 2012.05.14

(87)PCT国际申请的公布数据
WO2013/170429 ZH 2013.11.21

(73)专利权人 深圳华大基因股份有限公司
地址 518083 广东省深圳市盐田区洪安三街21号华大综合园7栋7层-14层

(72)发明人 陈盛培 葛会娟 李旭超 易贤 汪建 王俊 杨焕明 张秀清

(74)专利代理机构 北京清亦华知识产权代理事务所(普通合伙) 11201
代理人 李志东

(51)Int.Cl.
C12Q 1/68(2006.01)

(56)对比文件
CN 102317473 A,2012.01.11,说明书第77-161段.
CN 102127818 A,2011.07.20,说明书第8-37段.

审查员 蔡放

权利要求书3页 说明书14页 附图2页

(54)发明名称

确定胎儿基因组中预定区域碱基信息的方法、系统和计算机可读介质

(57)摘要

提供了确定胎儿基因组中预定区域碱基信息的方法、系统和计算机可读介质。其中,确定胎儿基因组中预定区域碱基信息的方法,包括下列步骤:针对胎儿基因组DNA样本,构建测序文库;对测序文库进行测序,以便获得胎儿的测序结果,该胎儿的测序结果由多个测序数据构成;基于胎儿的测序结果,结合胎儿遗传相关个体的遗传信息,根据隐马尔可夫模型,确定预定区域的碱基信息。

CN 104053789 B