[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

Pacific Biosciences of California EXHIBIT - 1046 (Part 3)
Pacific Biosciences of California v. The Chinese U. of Hong Kong - IPR2024-00028

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	chr7:82386121-82386138	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	chr5:13719018-13719028	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
rs104748 1.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	78381651783826	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	chr6:31023182-31023192	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.

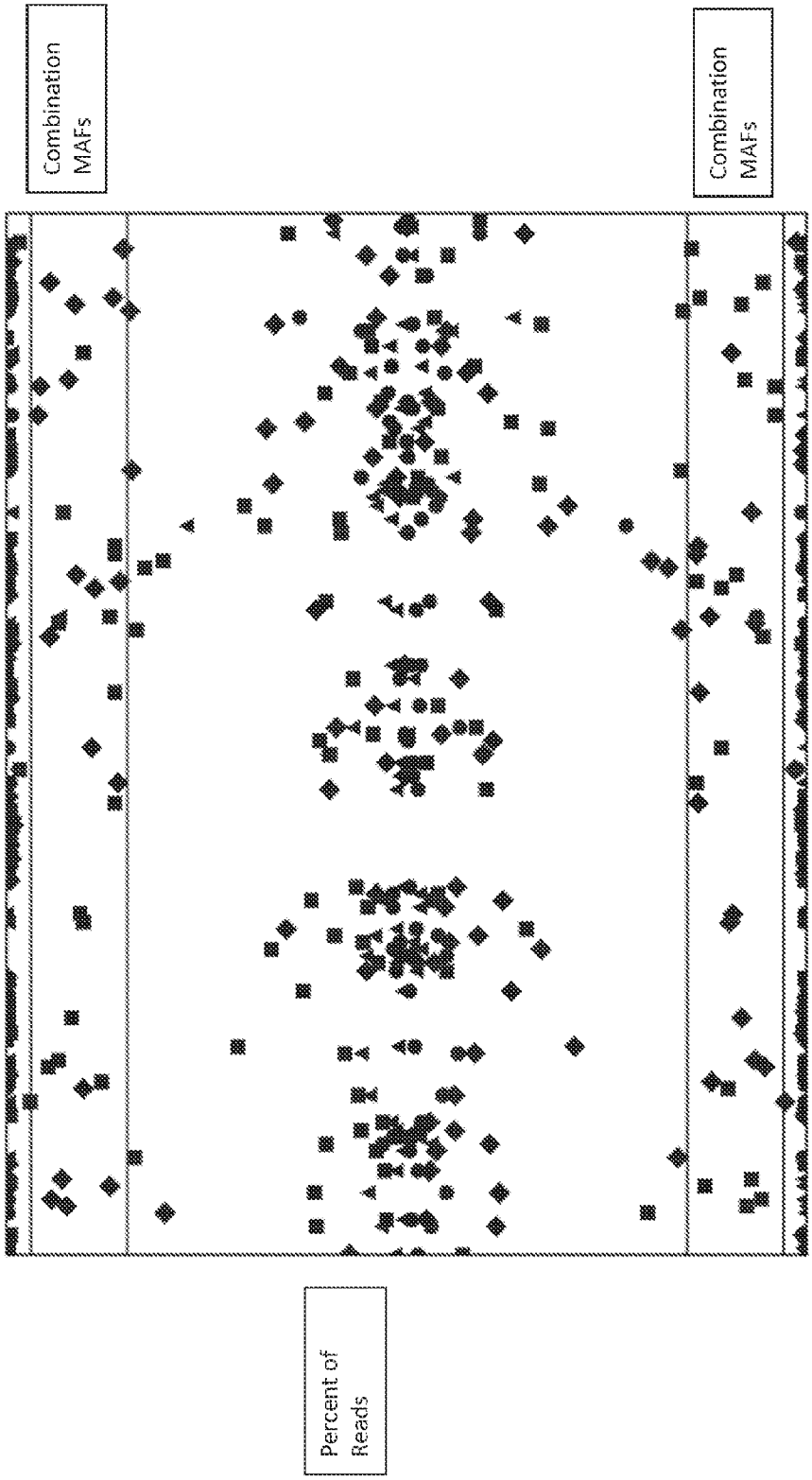


Figure 1

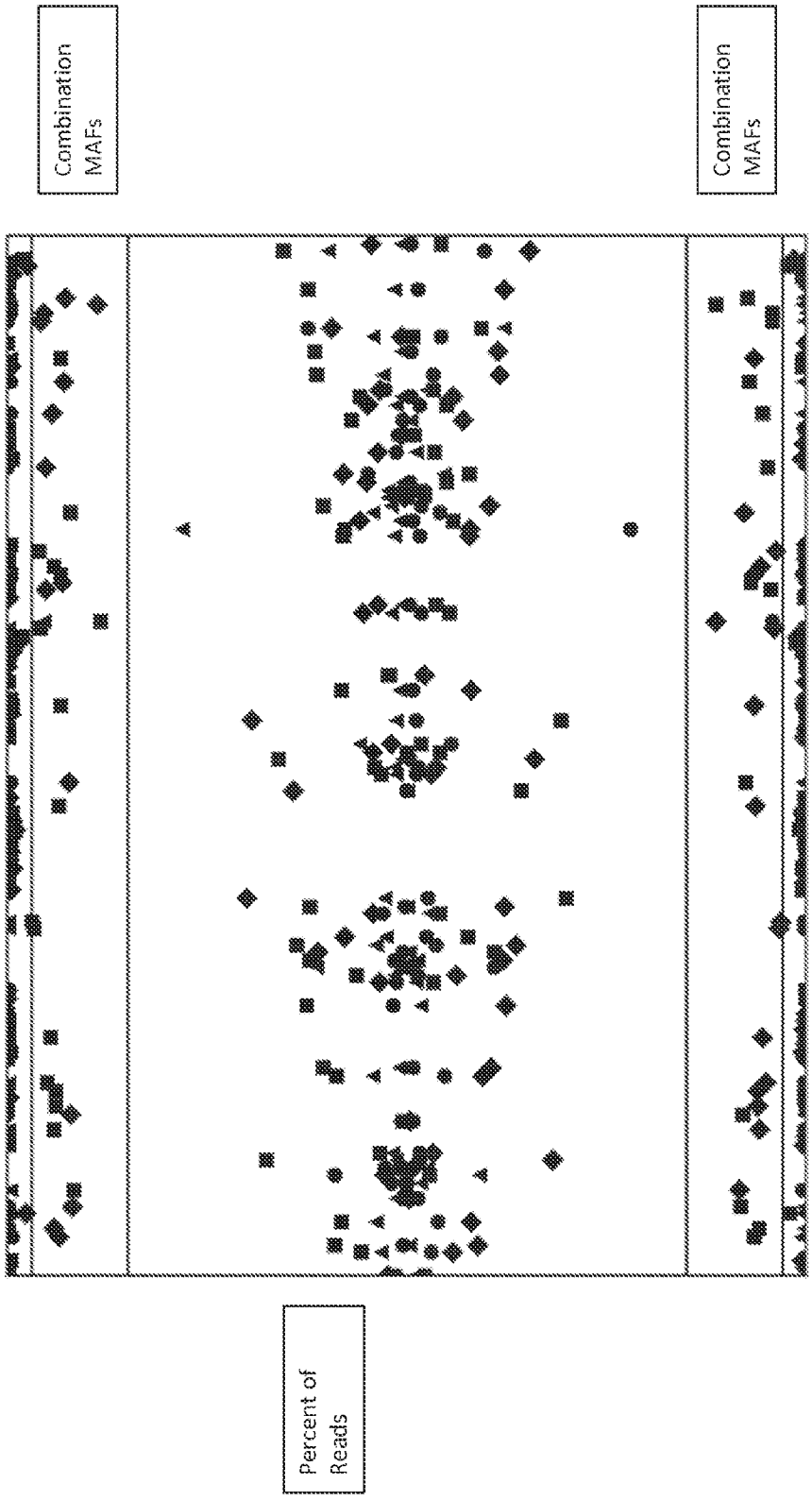


Figure 2

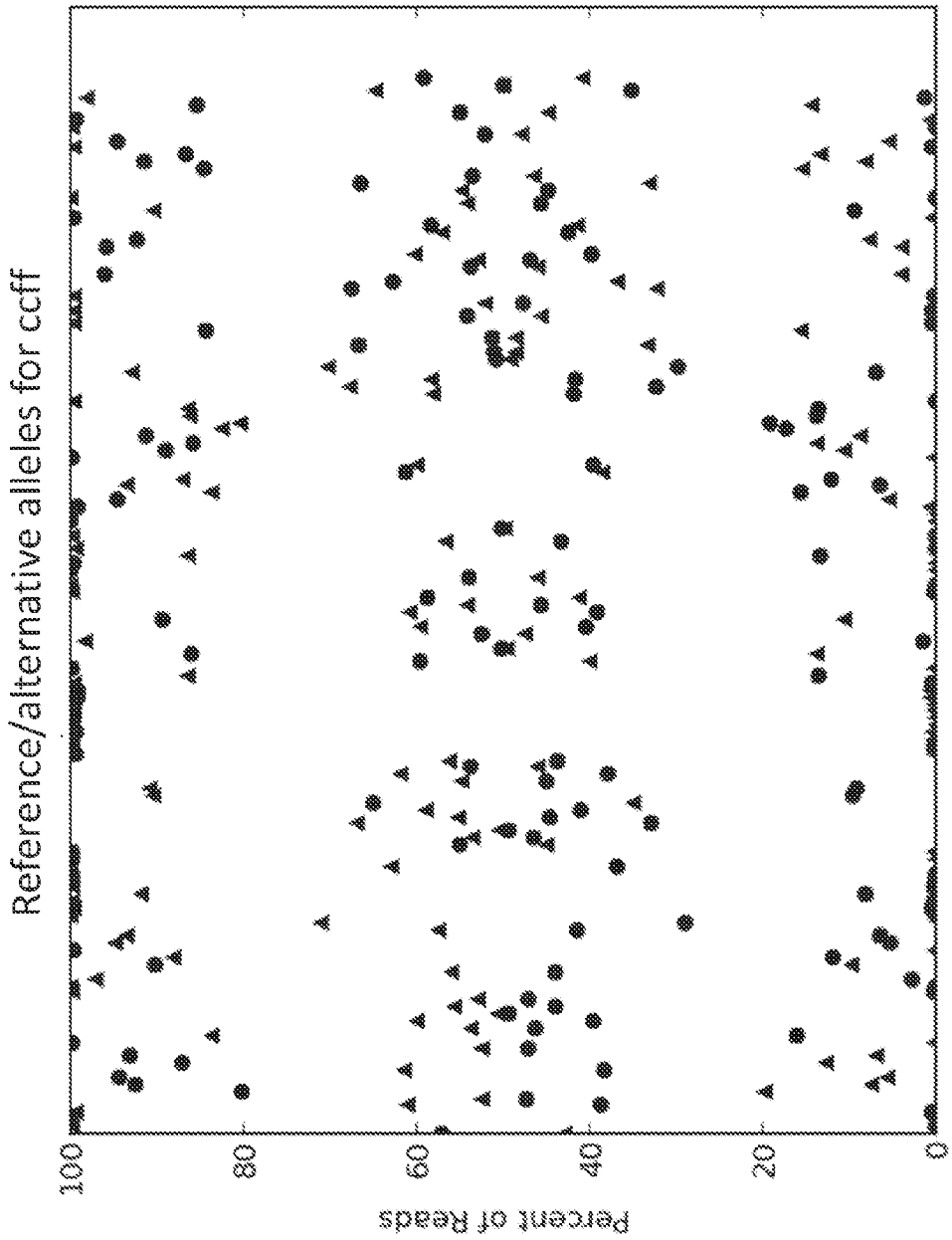

Reference/alternative alleles for ccff
Percent of Reads
100
80
60
40
20
0

Figure 3

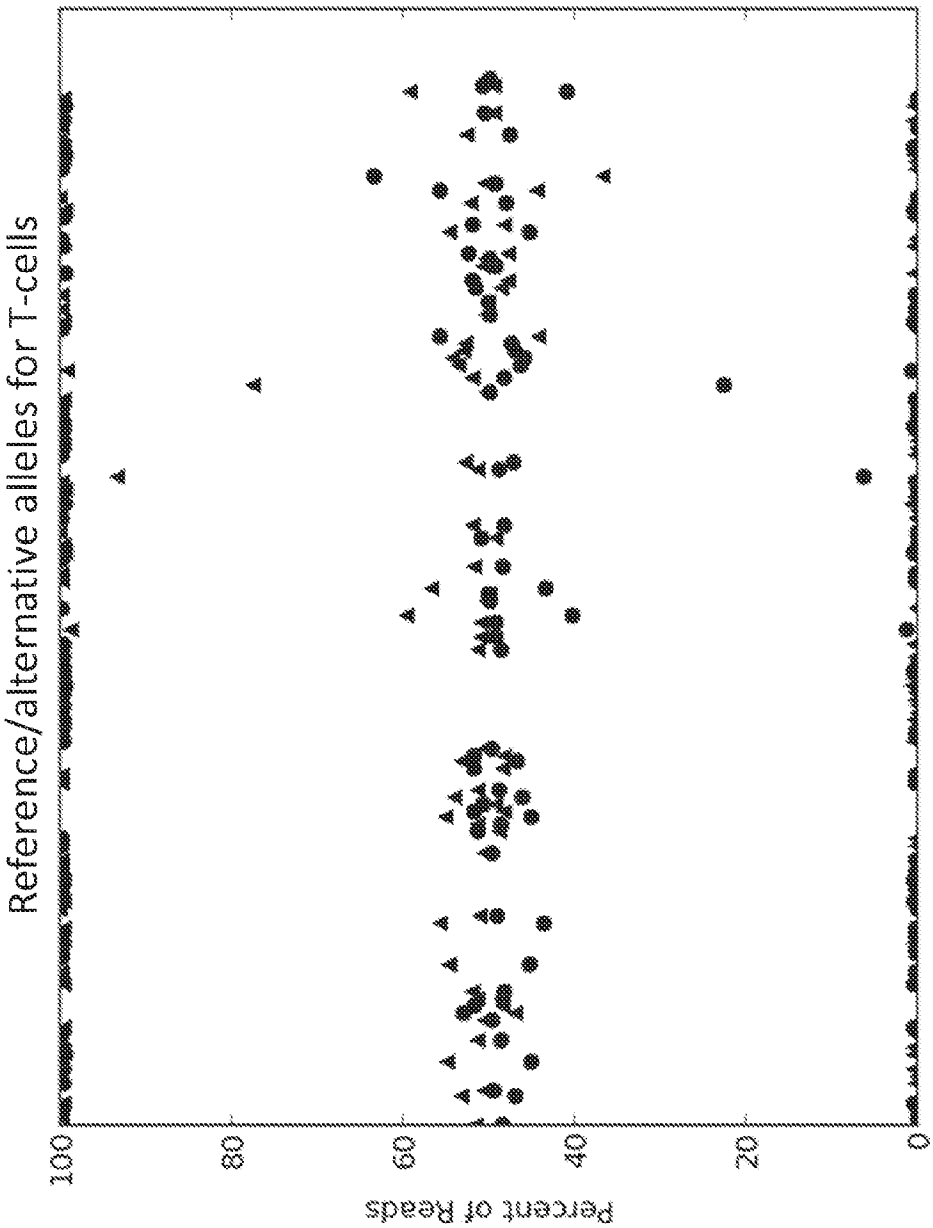

Reference/alternative alleles for T-cells
Percent of Reads
100
80
60
40
20
0

Figure 4

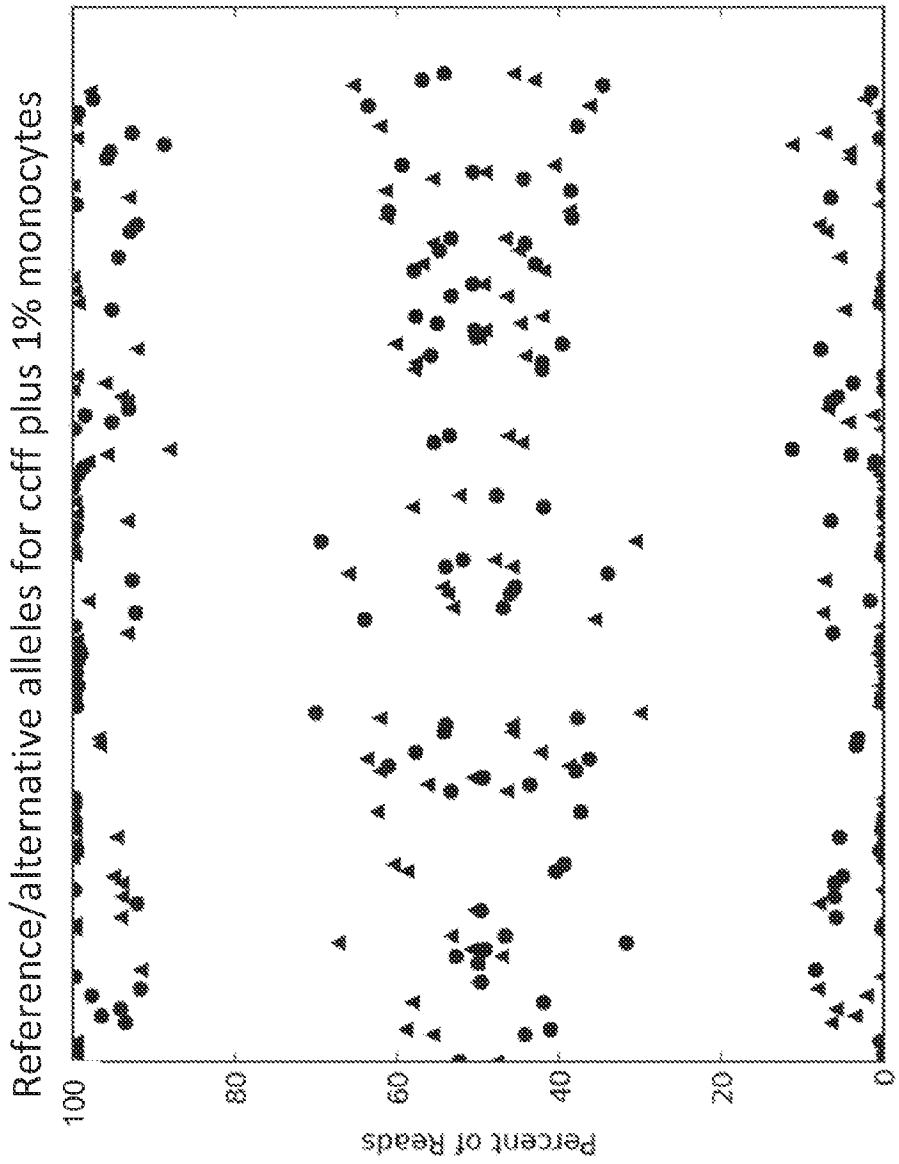

Reference/alternative alleles for ccff plus 1% monocytes
Percent of Reads
100
80
60
40
20
0

Figure 5

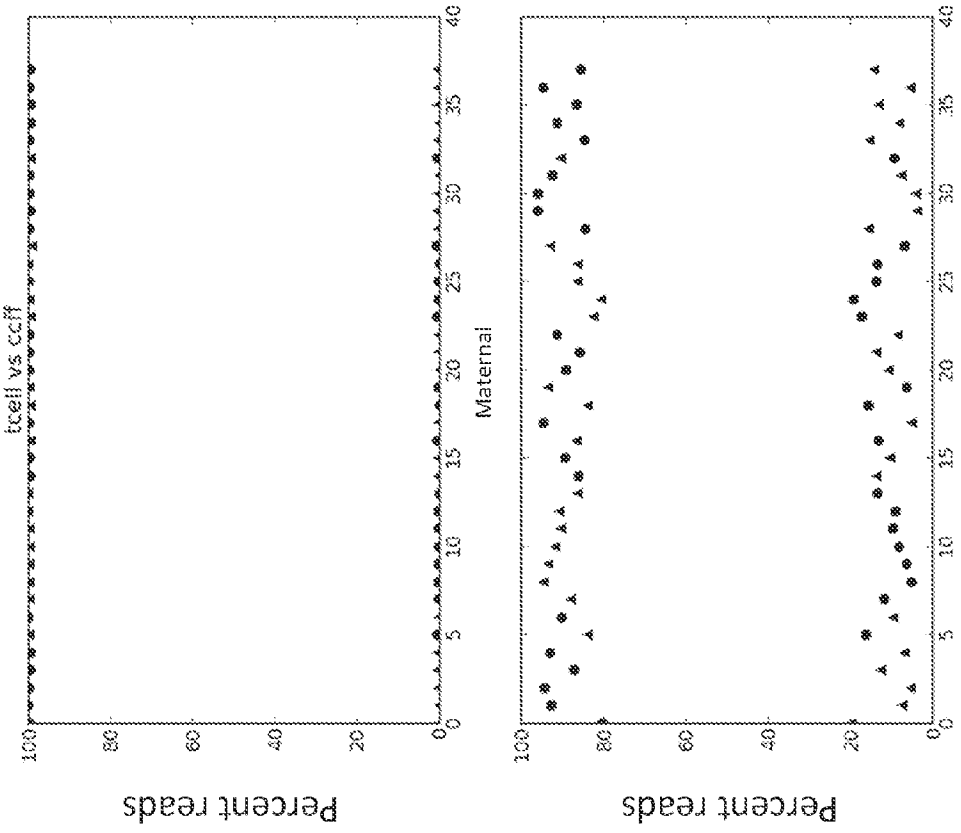

tcell vs ccff
Percent reads
Maternal
Percent reads
Combination

Figure 6

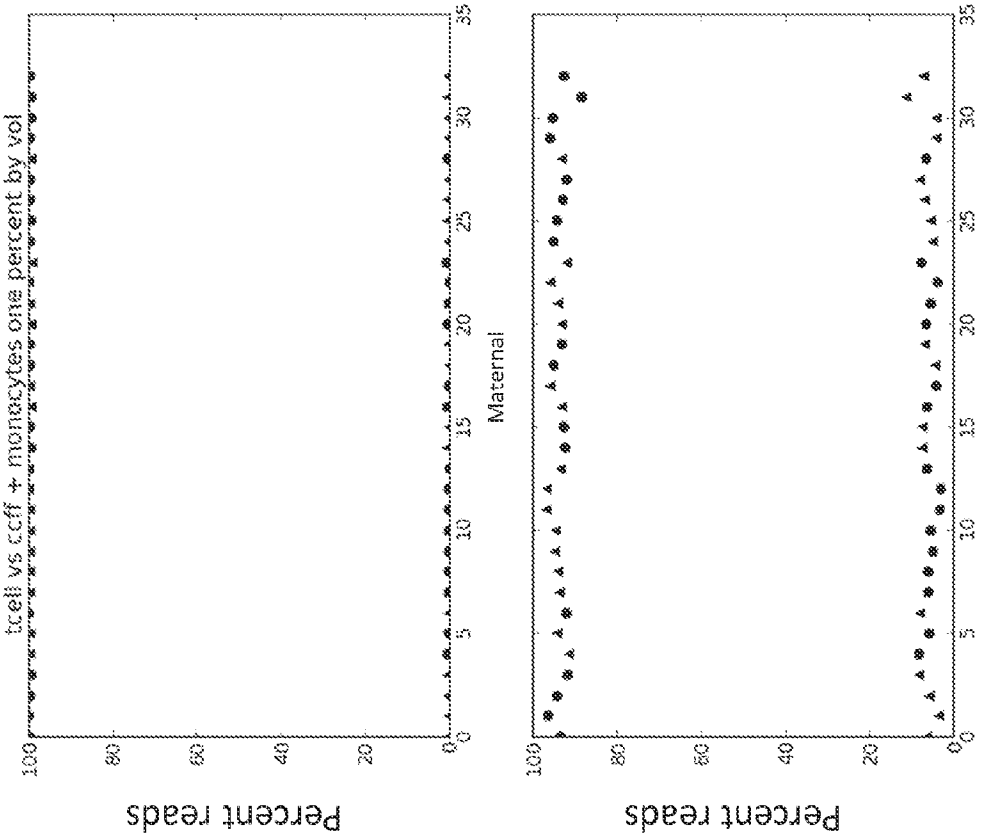

tcell vs ccff + monocytes one percent by vol
Percent reads
Maternal
Percent reads
Combination

Figure 7

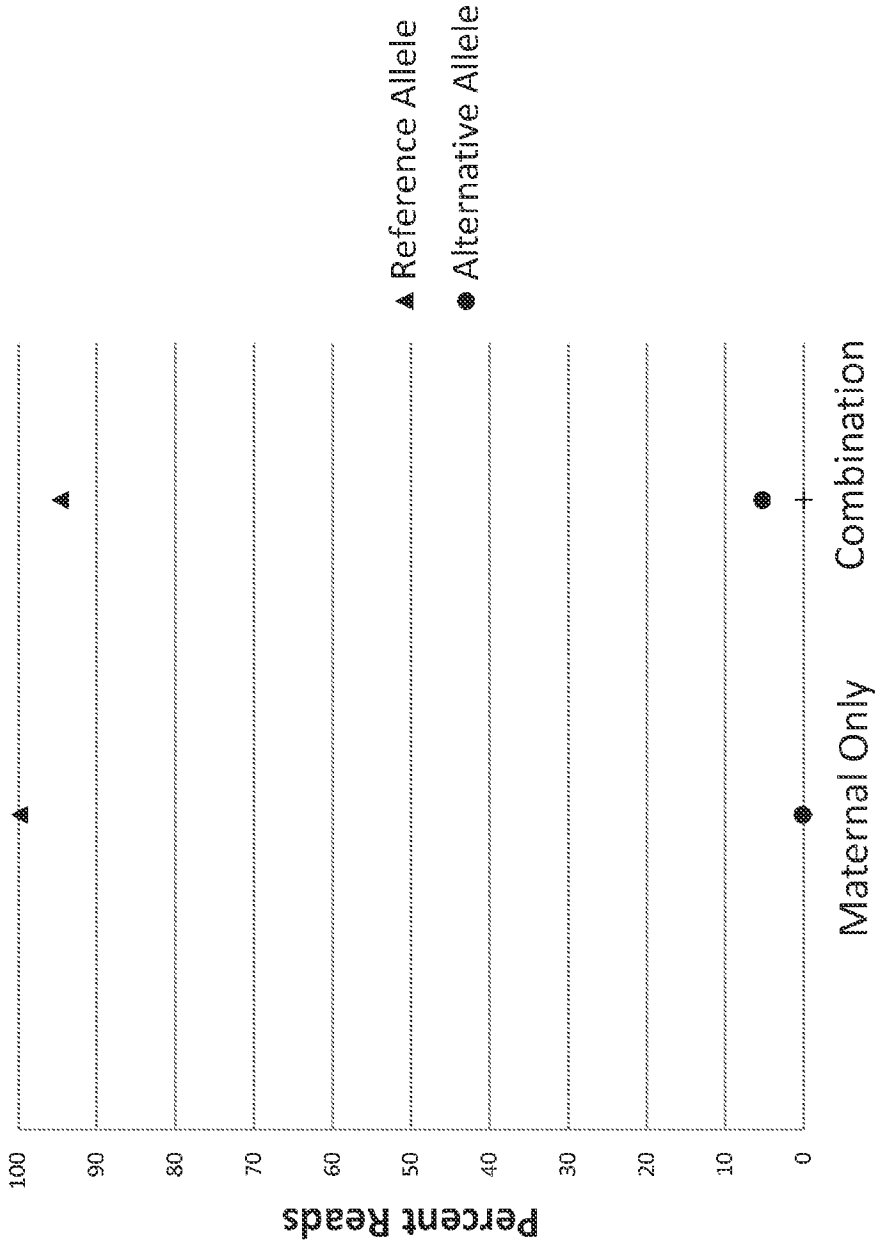

Reference Allele
Alternative Allele
Percent Reads
100
90
80
70
60
50
40
30
20
10
0
Maternal Only
Combination

Figure 8

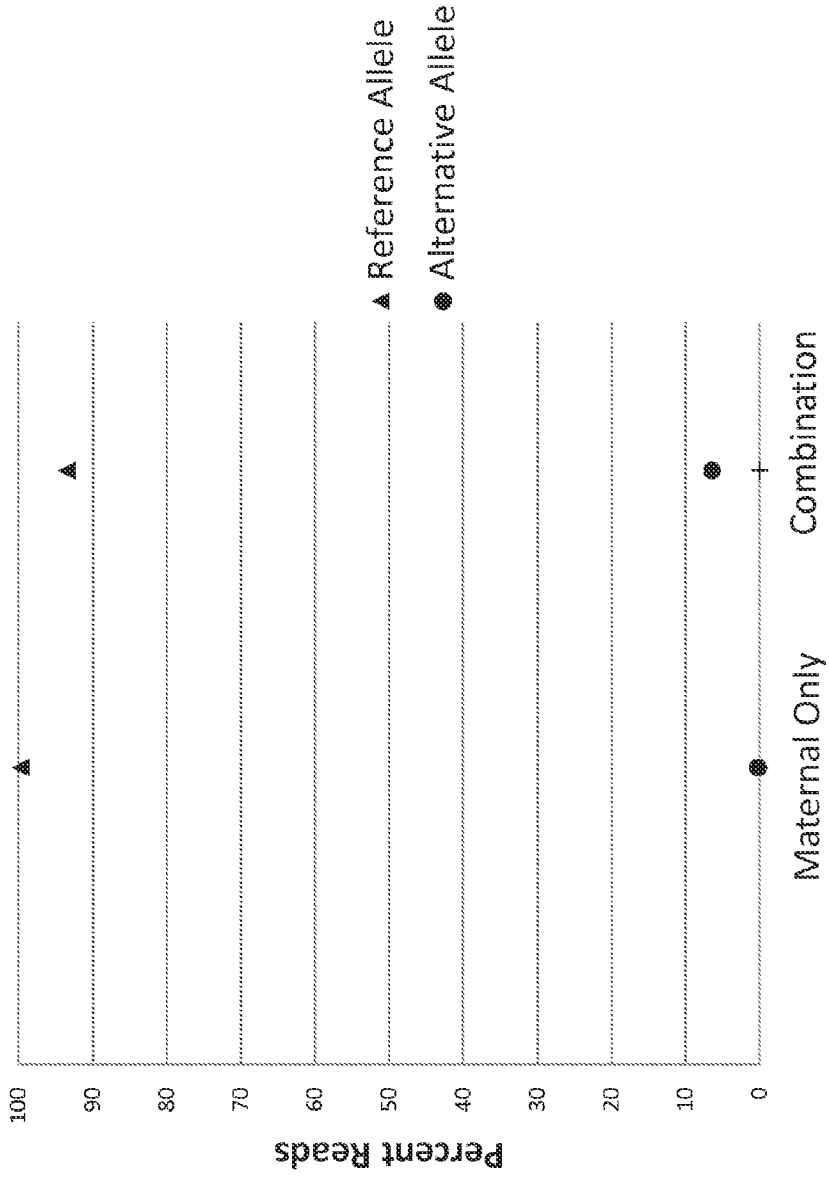

Reference Allele
Alternative Allele
Percent Reads
100
90
80
70
60
50
40
30
20
10
0
Maternal Only
Combination

Figure 9

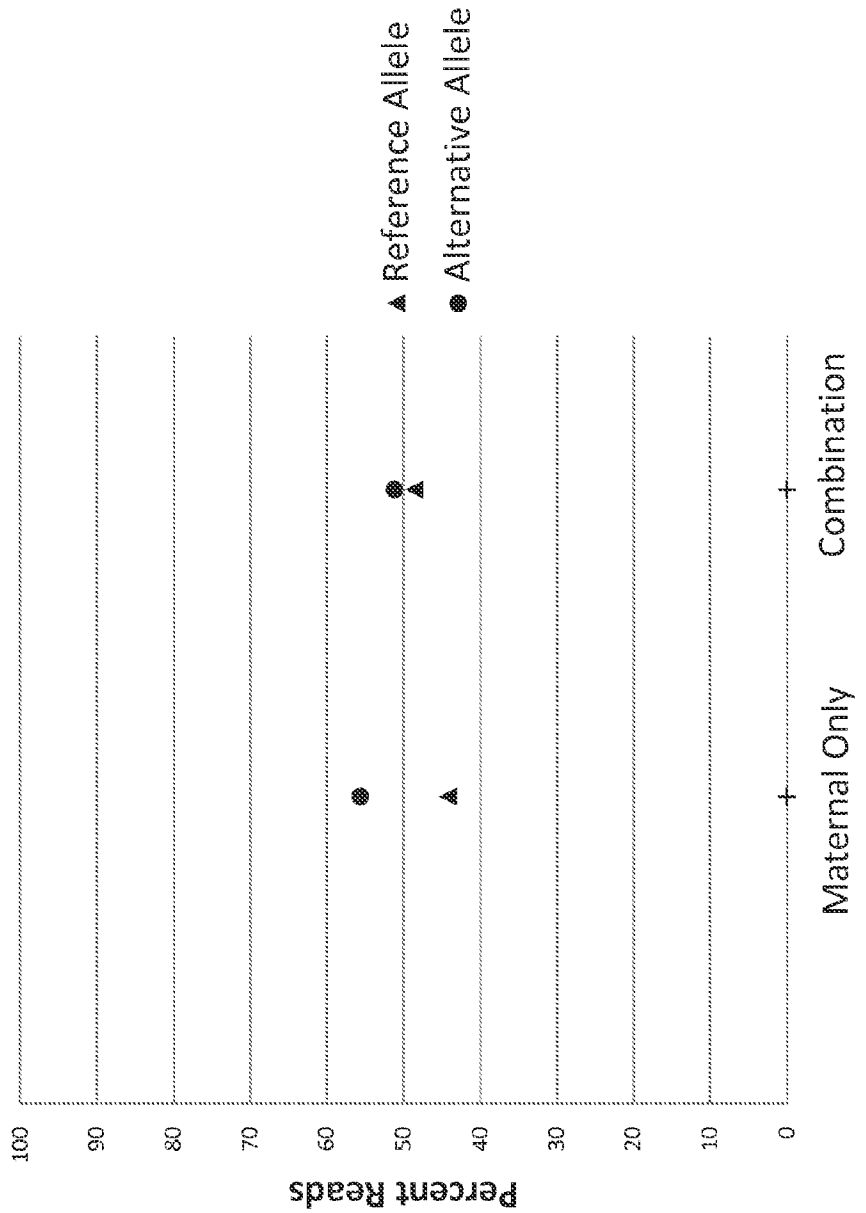

Reference Allele
Alternative Allele
Percent Reads
100
90
80
70
60
50
40
30
20
10
0
Maternal Only
Combination

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 Mail Stop PCT, Attn: ISA/US, Commissioner for Patents P.O. Box 1450, Alexandria, Virginia 22313-1450 Facsimile No. 571-273-3201	Authorized officer: 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.

UNITED STATES
PATENT AND TRADEMARK OFFICE

P.O. Box 1450
Alexandria, VA 22313 - 1450
www.uspto.gov

ELECTRONIC ACKNOWLEDGEMENT RECEIPT

APPLICATION #	RECEIPT DATE / TIME	ATTORNEY DOCKET #
17/891,899	**02/02/2023 12:46:58 PM ET**	**080015-028430US-1343271**

Title of Invention

DETERMINATION OF BASE MODIFICATIONS OF NUCLEIC ACIDS

Application Information

APPLICATION TYPE	Utility - Nonprovisional Application under 35 USC 111(a)	PATENT #	-
CONFIRMATION #	1331	FILED BY	Stephen Quast
PATENT CENTER #	61533929	FILING DATE	-
CUSTOMER #	20350	FIRST NAMED INVENTOR	Yuk-Ming Dennis Lo
CORRESPONDENCE ADDRESS	-	AUTHORIZED BY	David Hsu

Documents

TOTAL DOCUMENTS: 53

DOCUMENT	PAGES	DESCRIPTION	SIZE (KB)
IDS_080015-1343271.pdf	6	Information Disclosure Statement (IDS) Form (SB08)	1841 KB
Warning: This is not a USPTO supplied IDS fillable form. Data in the form cannot be automatically loaded to other USPTO systems.			
Transmittal_080015-1343271.pdf	2	Transmittal Letter	286 KB
FOR_AU_2002214811_B2.pdf	73	Foreign Reference	4115 KB
FOR_AU_2012380221_B2.pdf	34	Foreign Reference	1446 KB

FOR_CN_104053789_A.pdf	3	Foreign Reference	970 KB
FOR_WO_2005024053_A1.pdf	50	Foreign Reference	3219 KB
FOR_WO_2009013492_A1.pdf	115	Foreign Reference	6070 KB
FOR_WO_2010027484_A2.pdf	117	Foreign Reference	5927 KB
FOR_WO_2010068289_A2.pdf	103	Foreign Reference	5666 KB
FOR_WO_2012170429_A1.pdf	30	Foreign Reference	1725 KB
FOR_WO_2013170429_A1.pdf	30	Foreign Reference	3890 KB
FOR_WO_2013188846_A1.pdf	197	Foreign Reference	9077 KB
FOR_WO_2014153757_A1.pdf	32	Foreign Reference	4573 KB
FOR_WO_2017012544_A1.pdf	127	Foreign Reference	5843 KB
NPL_AU_2020323958_1ExmOA_05312021.pdf	3	Non Patent Literature	155 KB
NPL_AU_2022231719_1ExmOA_09282022.pdf	4	Non Patent Literature	185 KB
NPL_BEAULAURIER_2019.pdf	16	Non Patent Literature	1942 KB
NPL_BLOW_2016.pdf	28	Non Patent Literature	17722 KB

NPL_CLARK_2012.pdf	12	Non Patent Literature	9655 KB
NPL_CLARK_2013.pdf	10	Non Patent Literature	7656 KB
NPL_EID_2009.pdf	6	Non Patent Literature	920 KB
NPL_EP_20853612-8_EESR_10252021.pdf	15	Non Patent Literature	556 KB
NPL_FENG_2013.pdf	10	Non Patent Literature	6460 KB
NPL_FENG_2014.pdf	12	Non Patent Literature	11990 KB
NPL_FLUSBERG_2010.pdf	5	Non Patent Literature	6266 KB
NPL_GB_2102808-9_1ExmOA_04122021.pdf	6	Non Patent Literature	295 KB
NPL_GB_2102808-9_OA_06232021.pdf	8	Non Patent Literature	646 KB
NPL_GB_2102808-9_OA_09102021.pdf	4	Non Patent Literature	379 KB
NPL_GB_2103003-6_CSER_03292021.pdf	5	Non Patent Literature	437 KB
NPL_GB_2103010-1_CSER_04152021.pdf	4	Non Patent Literature	164 KB
NPL_GB_2201821-2_SR_02242022.pdf	4	Non Patent Literature	356 KB
NPL_GB_2210631-4_SR_11072022.pdf	4	Non Patent Literature	148 KB
NPL_GB_2210633-0_SEAR_08262022.pdf	4	Non Patent Literature	337 KB
NPL_GB_2216178-	4	Non Patent Literature	149 KB

0_SEAR_11152022.pdf			
NPL_GB2201811-3_CSER_03012022.pdf	4	Non Patent Literature	177 KB
NPL_HE_2017.pdf	2	Non Patent Literature	828 KB
NPL_JP_2021514525_OA_07262022.pdf	7	Non Patent Literature	4094 KB
NPL_KILE_2010.pdf	9	Non Patent Literature	2274 KB
NPL_LIU_Bisulfite_Free_Direct_2019.pdf	11	Non Patent Literature	8615 KB
NPL_LIU_Detection_of_DNA_2019.pdf	12	Non Patent Literature	16628 KB
NPL_Machine_Learning_2021.pdf	21	Non Patent Literature	1940 KB
NPL_NI_2019.pdf	10	Non Patent Literature	7586 KB
NPL_PCTCN2020109602_ISR_WO_11192020.pdf	11	Non Patent Literature	4833 KB
NPL_SCHADT_2013.pdf	13	Non Patent Literature	13638 KB
NPL_SG_11202101070Q_WO_10282021.pdf	9	Non Patent Literature	1540 KB
NPL_SUGAI_2017.pdf	10	Non Patent Literature	2967 KB
NPL_SUMMERER_2010.pdf	3	Non Patent Literature	2941 KB
NPL_SUZUKI_2016.pdf	9	Non Patent Literature	7599 KB
NPL_SUZUKI_2017.pdf	117	Non Patent Literature	18359 KB

NPL_US_16995607_NFOA_12302020.pdf	19	Non Patent Literature	2266 KB
NPL_US_17379544_NFOA_02042022.pdf	10	Non Patent Literature	361 KB
FOR_KR_101896167_B1.pdf	98	Foreign Reference	15044 KB
FOR_CN_104053789_B.pdf	21	Foreign Reference	21406 KB

Digest

DOCUMENT	MESSAGE DIGEST(SHA-512)
IDS_080015-1343271.pdf	69377A5ADC268C8C98FA967473BE38498F53C9666A0FA3863F21527BBB06D8B1D555A7F30D6D51B711843A688187500BB824A1ED1FC992D0474042F7567833CD
Transmittial_080015-1343271.pdf	EBE92C53353A6F0688F1FA6E6AB33689EC04F131BEB52A36559B8247FE232F31CBFBC51FEFE1CE1E4EFFE06F4A3BAD58E852C986F8D14B867EACE3D3D9AA638C
FOR_AU_2002214811_B2.pdf	BB5BD82939D30C42CBA2C8E760AE5563C6DD585CC03E3D8805FE81F22D3741572C102FCBAF0ACD99B9C4E79FF97BA7B5F8AF198695FF8B13E80B4B8565800248
FOR_AU_2012380221_B2.pdf	3432D07FC77B38FCC7ED2ABB174D039EBF8D6CBDBD7DF8E1FCF56FE81D42D8D408B1D95EF40FF53B3EB9B8B3916B4562F37A688411E532B9819AE3A804C4C794
FOR_CN_104053789_A.pdf	B0F72AC68BC680EE3E541F5933C68346E65E3274FCBEAAD0882A1120FC7DB7C3D80615DC0B200A434B463D94D13DC5A1F9995C60D3FA24A1286FE05C5CDD2B6B
FOR_WO_2005024053_A1.pdf	4B475956D7EF363E9EAC36DD2FA76E0B0B1D400CA76B828308AE9E9B4F8E651396DC2C822648B6C6F51ACCF21F8F3875DE8162A8E0E739680C2CE8DE894A5CF9

FOR_WO_2009013492_A1.pdf	2D31A5162570E1F075569D9227B9376692A2B47A604CD3B70BC6C9C2FB7E383B6F68742FCDF8A72C7B415FDAA02907C7A685F8854080B9A068DB7FCDB6BA03F7
FOR_WO_2010027484_A2.pdf	3F85AF638F68628F774A24ED5D4B63D82BD547817FA1FB46DB353A576D31A4FFC0632A97AE585ED43CB20625380DFCF69EEA510FE5E531E38BF8FA7D45DBD61F
FOR_WO_2010068289_A2.pdf	E99DACE237155FC231A954CE3A62C3F731B43006BB35682D86C978D0078D9764C07885F243212B6FC630309846179273389B9E7553E996E262F83810DD481F11
FOR_WO_2012170429_A1.pdf	D8503377C8764D5F56E90108A9D33A59DAB71AC3246CF6F8E237EA7057E030BB5510FAC4F2877F0F134D460C320E8566FB4D08007189217C4751B42BEE5F8B5E
FOR_WO_2013170429_A1.pdf	64DBA94B904E8B3E489FB77406A0A77424AE5F255B4FF6DBEDC33C7F519595F9EC0BEB3C2EF1408762FAE159647E59C6A03968554C98C15AC329B8DD363FD99C
FOR_WO_2013188846_A1.pdf	BC8C7596DE229BFA1CA4D1507924B9F0CA096C33E350261CF351E4B26D876E29BA8FBF10EC0ED073D0CF7E1F3CC995864ED49CDF7470C403C8CC67C23AD97660
FOR_WO_2014153757_A1.pdf	BBAB52A8215E6BCF5B616C8DA52A83193B3EB84FB533EABE81DE0E02A5539F8C3E855E011179E1710B2B1848A7FCDC3963D328E174E01140C844888EFF8F8917
FOR_WO_2017012544_A1.pdf	C001A3A0AE1450F5E08F268411848B761438E57F905EA43A2CF9EFC722C6E3C24B3FB258897C2DF88DEF73120FBE7F9E524FF39BD155AEB21985AE266203A274
NPL_AU_2020323958_1ExmOA_05312021.pdf	C1DA75351A918B499E76574E9749D7838118B0C3774CFC935BE4E2787C93808F75DAFC454E59FC0690BA17F0A8DCC436566C366329FFC322F21305E8B6E76785
NPL_AU_2022231719_1ExmOA_09282022.pdf	908762AB6AEAB34AB85DD3D7EAB0BDD834F5A59856BA7DD2E46D6CADBCA6A74E84324B27AD47F7DAC5BECCF754FCFB1BE9BB0CE7351992C59486774E4B343E33

NPL_BEAULAURIER_2019.pdf	EF65E0B813DD393320762000CF00F543735CF414C2663E2C36EF07A23E8BB3B4AA377EA33667A07129133AB43EFE06D714774D66AC93C029DFD53CA4A6069DF8
NPL_BLOW_2016.pdf	6F735E43AB863B91CCDD2BA47EE5F60DF2A1F4BA8688EFC0D7FD3B43F6CFF33A666AA3298C75DF97FF8D8FE58EBE0E623A8341CE0C66B476CC36CC810D135DE1
NPL_CLARK_2012.pdf	DD040B2423E651819F020F8958D4FC65AC38EE793B2409807911A642661B972666D1C6C08DACAFEE5886FE57D4C37CCBCC4B4C226B470BEBC5885A142623EC83C
NPL_CLARK_2013.pdf	80BF1EB1DB9C88B0304B92211FBDD26C0D6C4703618660BC6E83A73BC1AB023678993BAC03C5D9788A5E2F38E1F792893926BD1F0524959B0996EE323D2900E7
NPL_EID_2009.pdf	34BDA4B0E2F6E13363278F19FC2E6C373BBA9F1934DA13A2B0CEB0407759FE12F34E1F3320935D8D0943E0ABD30C2C15F6280A13510270119EE752E08CFC1741
NPL_EP_20853612-8_EESR_10252021.pdf	ED1239BC8B290A0EA88A6CDCA61A37E01384C77D4F5D68A4534FF7DB435AEA17768A8612C1FD4386787A5EED8AEAD10AC19A38052D12725F658B82339F428606
NPL_FENG_2013.pdf	DB891DCA2C9D65811BA428A1CAAD7F36EA950000A54EF00423D6B3540E88A3621D8FBDE3C63C29863FD4FD036C18A99D6686F8255C6DE5BF5FF6E74FB3EC6520
NPL_FENG_2014.pdf	3167107F6B63357198553AFCC2A462A4BACE5B338F4DD9C6FB41A7D7831B2530ADB865A87F3A082A49A32FA1D22808B300299D644106FC5C60DBAACF2428F16C
NPL_FLUSBERG_2010.pdf	DBDA87CA229F07681C83F713FD355BF3F723695405DBF76BE91399F3399F17BE7D64AD7ECAF371A1ACECE615F6E80636DA4D5C83B5F88745F7AB326D9AD8CB2E
NPL_GB_2102808-9_1ExmOA_04122021.pdf	044DCF8FE638E8D30FFC5BD1D4A9C088C38CBC88231D1939A0FD65983207464232E8F92D40E6D38F00956845A926C52F09B1758C9E7F87E91868D38C9F4CC539

NPL_GB_2102808-9_OA_06232021.pdf	4CAF89D129BEC175C7AD1B7B376916C7671DB2214CEACDFE5B969324C630F078B2672A6BA4DCFAC1E72FCC80CB6905625CE072990D92F6A9A4C7BD559A0AE62F
NPL_GB_2102808-9_OA_09102021.pdf	C18A8BDD37E83781B603FB9C59DF13F76B4498D34EFB4315C06BF16372A7E58CAC6B6B4727D000E223E040D35A5014685C6A0A5765F0F965B4739CF3748BBB2C
NPL_GB_2103003-6_CSER_03292021.pdf	C1B494E8473575EF70CF65A5758F61BE796E17A31B9CB537A84850AA690EA2ECB14E09859AC0EF9F42651D36437AB1BB7086B475A93E86129598CA70F267026F
NPL_GB_2103010-1_CSER_04152021.pdf	34A487E57CDD6D958C596079A847062611C4299D679F742C69188725E65D0ED73A8F806670BA613C20CBE40CFC781849546CCAFCD9753E717746D2C7E20AB9AD
NPL_GB_2201821-2_SR_02242022.pdf	AD6443046F73A6921C121B2C822C3D1B1FFC8766337E9511208353CC3FFACCEB59A65CAB53D5F7FA99C515B205087C5B1DC3FD677FB51C045B05736507FEAC14
NPL_GB_2210631-4_SR_11072022.pdf	B2459EDC993A13AAE33A5C1D4AA7BF692AA3315516817857E7DE4ECE621684D9A78780963D9E5DE5C7C5DA611A548827E273917A9783D9A63434B5A237F739ED
NPL_GB_2210633-0_SEAR_08262022.pdf	D1FFEA9FF6117B029F3221923B86F47878A12BB182255C811044DE65A87465A833DC46C9581212949031 4C058E543D53831262428509A9A905945ABA6D58BC1B
NPL_GB_2216178-0_SEAR_11152022.pdf	9E3B14A1485B6B62266F5857F827868FB625DC81F6493451C057969D7BECAC4B119B60458BCFC40251763AC894D81795F7AD25A7679AEF2A1AACE9A36B7839F8
NPL_GB2201811-3_CSER_03012022.pdf	C07F861F197D8FF98BC7BB8DB71BEF5AAABC118681DA7A7F22E562F1819AD7CC4B2F491CC7F1B4A9EBB3BD2EA955B7BCC3611CF65B8B6E97D9ED94F9D980F97C
NPL_HE_2017.pdf	1CB13A4B35E0A958291B5F189887FB30893CFEE9F066579E0C5221C00A1C366282BF83F4605928DF4CC681A1E05AE4E4360D115EB06D642B9F1C062C634B0AA1

NPL_JP_2021514525_OA_07262022.pdf	8561A3CC8A123513DC2C7AE4C86474726208B13454EF3505B832B1635504BB2F66B19DC60F9DE18D37160F0D6511F55B0879DC82C0D12AF591902ED5DC54EE53
NPL_KILE_2010.pdf	38DECE0634674931CD84F80C08BC11C56567464ED8217F75D9400AAD12C3544F38AC605B8FF923B572CEC5A3E657B2BF0B7C5ED118FA1A99E983B3C9265EB342
NPL_LIU_Bisulfite_Free_Direct_2019.pdf	0BC8A24C84C65E808E90FA3EA2DE393CCBC721FBE9902FC0D776F99BA47F31692BB330A672CB2985370919098FA6F1BD4C4DE4AF3E4F77EF14B91766FF4731BB
NPL_LIU_Detection_of_DNA_2019.pdf	42A2C1D92B5EA53B46CFE1FB46279642783CD5B617E8F3409ADCF77706B838E01C3579A56F5F1EDBD6594306DF8B11D5778DDCF90E2A43D618B99DB6B2C3B259
NPL_Machine_Learning_2021.pdf	B3F3148CFFE062095D2D68811B15D380813967AAB086BC08339EA97EBE108B762FE0B805A0C2A8A3450C5FBC2A99CB0069685A4F19B29728920E0AA595DCF19F
NPL_NI_2019.pdf	B8C9EC53CB7B6E9778211A716846B87846FBE50AEF50D55A0E5EF50EFD0A73E59E96625D40892A6D4A96BFFA9B78E070B3332A9F102F9D498DA53DCAA45687EE
NPL_PCTCN2020109602_ISR_WO_11192020.pdf	91094A9AAF3A4F82B716E5AED16510D6DD3CBF0BDE428470C057E76C9E9BAD240D300122D4F747C12C46CE791309DF9D7A24FB6E78F1E2BA0ADAED199EC8A070
NPL_SCHADT_2013.pdf	90EE6AC6AD9B8E557F9697932A437BC5D037FCCDC30182BF4BE4CBCE68088CC45AD0CD7D0471DDE7ACA3F9721397C9D683ACA46AE15483DCA5B4E3A9776C9024
NPL_SG_11202101070Q_WO_10282021.pdf	425B1AA46D0BA5BCE0CA2A65D05ACF3F14F3853CCD3C897822601136DC580496F47D97DADBBB996DA24DAA7E417AA7B460FA681623E1BFF93EEAF83480979581
NPL_SUGAI_2017.pdf	38A0C6B78449818 3C05778D3019797DEB9A747BCFDBDF8531031541E329AA6C8AAD09C0934D89642746BBE3EB2193C7789BA029C3FAA8FE35BE427477B49573D

NPL_SUMMERER_2010.pdf	ED271AF9790C5F28A93E1CEA35290F2A6CE9197BC13AFCBD539DF7EF1EE3294192E19E4F8BF89117743769184 0F8D7E545686A733EB2EF618A0F8A87744483A2
NPL_SUZUKI_2016.pdf	C38FE082108EF315300AB0C1D2DE8FDC15B666785ACB29F46AD4660414A14C2A25F1CF7B85E5732FCAA9D1CC3D80F329363193F708A8302EF0BD2F58F377296A
NPL_SUZUKI_2017.pdf	B62AD5D5E28FB94645C5CD8504D591A8020B06C66FD186357AA7B1930113C5942F0A50DF5637A8FF5AF1BB42CE689FFAE8C5ACC0A1AD5C6F9AB5B875B9346E33
NPL_US_16995607_NFOA_12302020.pdf	59E4B5BDBB255F123FF71A2A23012939BF419568EF195F399823B31FD31734DD7AD75961418410AE628F67662476 1C5588D5DB7715098282CF24215C82125078
NPL_US_17379544_NFOA_02042022.pdf	245D06E6C52F309261B34B5119E85F589A889E0BE8DCD96CE413028D037E63B3070AFBF82C74B905A02EC12A081E6735A2FB827FC6DE5AF6650BAAAAAA6387B4
FOR_KR_101896167_B1.pdf	E3BB2927767F92741238A549082FCC46229D5C7FFD002EA6D8B8BF04ADC6F14DA923C00B360BA9CDBFBC052159D2B4613CBF2A5434707EF593E3C2B8BDCACE3F
FOR_CN_104053789_B.pdf	7D526AA430A68D40F6ABFD1F86135D3A59C4C43CE16B2AD412F638D110D7BC2602C99BAEDDA2281698E4CC417242184FF9CAB6F5C21A9A793C6C68E3373938BE

This Acknowledgement Receipt evidences receipt on the noted date by the USPTO of the indicated documents, characterized by the applicant, and including page counts, where applicable. It serves as evidence of receipt similar to a Post Card, as described in MPEP 503.

New Applications Under 35 U.S.C. 111

If a new application is being filed and the application includes the necessary components for filing date (see 37 CFR 1.53(b)-(d) and MPEP 506), a Filing Receipt (37 CFR 1.54) will be issued in due course and the date shown on this Acknowledgement Receipt will establish the filing date of the application

National Stage of an International Application under 35 U.S.C. 371

If a timely submission to enter the national stage of an international application is compliant with the conditions of 35 U.S.C. 371 and other applicable requirements a Form PCT/DO/EO/903 indicating acceptance of the application as a national stage submission under 35 U.S.C. 371 will be issued in addition to the Filing Receipt, in due course.

New International Application Filed with the USPTO as a Receiving Office

If a new international application is being filed and the international application includes the necessary components for an

international filing date (see PCT Article 11 and MPEP 1810), a Notification of the International Application Number and of the International Filing Date (Form PCT/RO/105) will be issued in due course, subject to prescriptions concerning national security, and the date shown on this Acknowledgement Receipt will establish the international filing date of the application.

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

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
11 March 2010 (11.03.2010)

PCT

(10) International Publication Number
WO 2010/027484 A2

(51) International Patent Classification:
C12Q 1/68 (2006.01) *C12Q 1/25* (2006.01)
C12N 15/11 (2006.01)

(21) International Application Number:
PCT/US2009/004993

(22) International Filing Date:
4 September 2009 (04.09.2009)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
61/094,843 5 September 2008 (05.09.2008) US
12/370,472 12 February 2009 (12.02.2009) 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)*: **PATEL, Pranav** [IN/US]; 34203 Oneil Terrace, Fremont, CA 94555 (US). **KORLACH, Jonas** [DE/US]; 36100 Toulouse Street, Newark, CA 94560 (US). **BIBILLO, Arkadiusz** [PL/US]; 10100 Torre Ave, Apt. 144, Cupertino, CA 95014 (US). **BJORNSON, Keith** [US/US]; 8098 Peachtree Avenue, Newark, CA 94560 (US). **HANES, Jeremiah** [US/US]; 3243 Spring Street, Redwood City, CA 94063 (US).

(74) Agents: **LITTLEPAGE, Paul** 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: ENGINEERING POLYMERASES AND REACTION CONDITIONS FOR MODIFIED INCORPORATION PROPERTIES

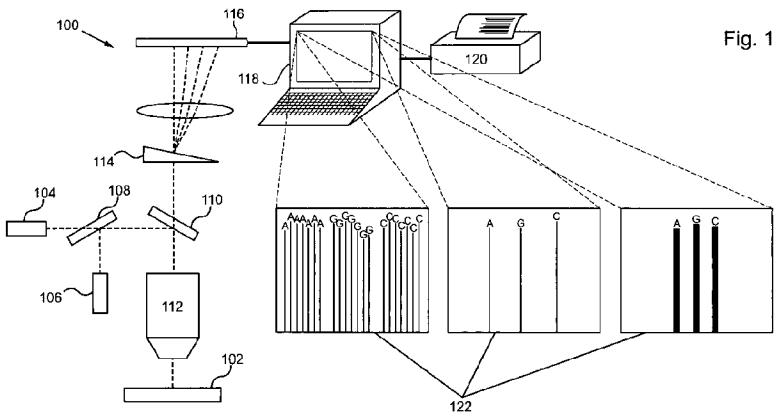


Fig. 1

(57) Abstract: Provided are methods for enhanced sequencing of nucleic acid templates. Also provided are reaction conditions that increase branching fractions during polymerization reactions. Also provided are compositions comprising modified recombinant polymerases that exhibit branching fractions that are higher than the branching fractions of the polymerases from which they were derived. Provided are compositions comprising modified recombinant polymerases that exhibit delayed translocation relative to the polymerases from which they were derived. Also provided are compositions comprising modified recombinant polymerases that exhibit increased nucleotide or nucleotide analog residence time at an active site of the polymerase. Provided are methods for generating polymerases with the aforementioned phenotypes and methods of using such polymerases to sequence a DNA template or make a DNA. Also provided are methods and nucleic acid sequencing systems for determining which labeled nucleotide is incorporated at a site during a template-dependent polymerization reaction.

WO 2010/027484 A2

ENGINEERING POLYMERASES AND REACTION CONDITIONS FOR MODIFIED INCORPORATION PROPERTIES

CROSS-REFERENCE TO RELATED APPLICATIONS

[0001] This application claims priority to Provisional U.S. Patent Application No. 61/094,843, filed September 5, 2008, and U.S. Patent Application No. 12/370,472, filed February 12, 2009, the full disclosures of which are incorporated herein by reference in their entirety for all purposes.

FIELD OF THE INVENTION

[0002] The invention is in the field of nucleic acid sequencing, for example, single molecule sequencing.

BACKGROUND OF THE INVENTION

[0003] High throughput sequencing has become a central tool in the field of biotechnology and is revolutionizing personalized medicine. Many diseases and/or disorders are genetic in origin. Acquiring the genomic sequence of individual patients in a comprehensive, rapid and cost-effective manner enhances the ability of medical professionals to diagnose diseases or identify predispositions to diseases or other genetic-based disorders. Genomic sequence information also enhances the treatment of diseases by providing doctors with information regarding the efficacy of a given therapy for a particular individual.

[0004] One approach aimed at efficiently obtaining the complete genomic sequence of an organism is sequencing by incorporation, where the identity of the sequence of nucleotides in a template nucleic acid polymer is determined by identifying each complementary base that is added to a nascent strand being synthesized against the template sequence, as such bases are added. While detection of added bases may be a result of detecting a byproduct of the synthesis or extension reaction, e.g., detecting released pyrophosphate, in many systems and processes, added bases are labeled with fluorescent dyes that permit their detection. By uniquely labeling each base with a distinguishable fluorescent dye, one attaches a distinctive

detectable characteristic to each dye that is incorporated, and as a result provides a basis for identification of an incorporated base, and by extension, its complementary base upon the template sequence.

[0005] During sequencing by incorporation, nucleotide (or nucleotide analog) incorporation events are detected in real-time as the bases are incorporated into the extension product. This can be accomplished by immobilizing the complex within an optically confined space or otherwise resolved as an individual molecular complex. Some sequencing by incorporation methods employ nucleotide analogs that include fluorescent labels coupled to the polyphosphate chain of the analog, which are then exposed to the complex. Upon incorporation, the nucleotide – along with its fluorescent label – is retained by the complex for a time and in a manner that permits the detection of a signal "pulse" from the fluorescent label at the incorporation site. Upon completion of incorporation, all but the alpha phosphate group of the nucleotide is cleaved away, liberating the label from retention by the complex, and diffusing the signal from that label.

[0006] Thus, during an incorporation event, a complementary nucleotide analog, including its fluorescent label, is effectively "immobilized" for a time at the incorporation site, and the fluorescent label is subsequently released and diffuses away when incorporation is completed. Detecting the localized "pulses" of fluorescent tags immobilized at the incorporation site, and distinguishing those pulses from a variety of other signals and background noise, allows bases to be called in real-time as they are incorporated. Further details regarding base calling during sequencing by incorporation methods are found in Tomaney et al. PCT Application Serial No. PCT/US2008/065996 METHODS AND PROCESSES FOR CALLING BASES IN SEQUENCING BY INCORPORATION METHODS, incorporated herein by reference in its entirety for all purposes.

[0007] Current real-time sequencing by incorporation methods may exhibit sub-optimal reliability and accuracy due to missed signal pulses that contribute as errors in sequencing reads. Missed pulses derive from, e.g., insufficient residence time of the analogs at an active site of the polymerase or unlabeled or broken-fluorophore nucleotide analogs. Compositions

and methods for improving the reliability and accuracy of sequencing by incorporation are desirable.

SUMMARY OF THE INVENTION

[0008] Altered reaction conditions and modified DNA polymerases can find use in such applications as, e.g., single-molecule sequencing (SMS), genotyping analyses such as SNP genotyping using single-base extension methods, and real-time monitoring of amplification, e.g., real time PCR. The invention provides methods of sequencing a nucleic acid template, which methods utilize signal pulses or signatures from branch fraction nonincorporation events (and, optionally, actual nucleotide incorporation events) to determine which nucleotide is incorporated at a particular site/position of the template nucleic acid. The invention further provides methods that modulate (e.g., increase) the branching rate of a polymerization reaction to facilitate identifying which nucleotide is incorporated at a particular site. A nucleic acid sequencing system that detects and utilizes signal pulses or signatures from branch fraction nonincorporation events to determine the sequence of a template nucleic acid is also provided by the invention. The invention further provides compositions that include modified recombinant polymerases that exhibit properties, e.g., increased branching fraction, delayed translocation or increased nucleotide or nucleotide analog residence time, which can be particularly desirable for these applications. These improved polymerase properties can facilitate readout accuracy. In addition, the invention provides methods of generating the modified polymerases of the invention and methods in which such polymerases can be used to e.g., sequence a DNA template and/or make a DNA.

[0009] In one aspect, the invention provides methods for determining which labeled nucleotide is incorporated at a particular site during a template dependent polymerization reaction. The methods include 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. The methods additionally include 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. The methods optionally comprise 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. Optionally, the polymerization reaction is a high branch fraction polymerization reaction, where the branch fraction is optionally 70% or more, 80% or more, or 90% or more.

[0010] The methods described above optionally include at least one species of metal ion, which metal ion increases the frequency of branch fraction nonincorporation events in the reaction. Example metal ions include: Mg^{++}, Mn^{++}, Zn^{++}, Co^{++}, Ca^{++}, Fe^{++}, Cr^{++} and Sr^{++}. The methods described above optionally comprise both Mg^{++} and Mn^{++}, e.g., where the concentration of Mg^{++} is higher than the concentration of Mn^{++}.

[0011] The methods described above optionally include 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. The modified recombinant polymerase can optionally exhibit a higher branching fraction as compared to a corresponding wild-type polymerase, or an increased exonuclease rate that is about 10% to 50% as compared to its polymerization rate.

[0012] The methods described above optionally include branch fraction nonincorporation events that comprise iterative sampling of labeled unincorporatable nucleotide analogs, optionally including actual nucleotide incorporation events that comprise incorporation of unlabeled nucleotides. Actual nucleotide incorporation events optionally include incorporation of nucleotides that are differentially labeled as compared to the unincorporatable nucleotides. In one example, labeled unincorporatable nucleotide analogs comprise a link between an alpha and beta phosphate group that is not hydrolyzable by a polymerase enzyme.

[0013] The methods described above optionally include a polymerase enzyme, polymerase reaction conditions, and/or polymerase reaction substrates that are selected such that the polymerization reaction exhibits two kinetically observable steps within an observable phase of the polymerase reaction. The two kinetically observable steps are optionally steps which proceed in a bright phase or a dark phase, and the polymerase enzyme optionally

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. Optionally, the polymerase reaction conditions can include, e.g., a selected metal cofactor concentration, a selected pH, a selected temperature, an enzyme activity modulator, D_2O, an organic solvent, and a buffer.

[0014] The methods described above optionally comprise branch fraction nonincorporation events that comprise noncognate branch fraction nonincorporation events. Branch fraction signal pulses or signatures are optionally generated from noncognate branching events of a nucleotide analog, e.g., guanine and thymine.

[0015] The reaction of the methods described above is optionally reacted in a DNA sequencing system, where the DNA sequencing system optionally comprises a zero mode waveguide or nanohole.

[0016] Assigning the nucleotide in the methods described above optionally 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.

[0017] The methods described above optionally comprise 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.

[0018] In another aspect, the invention provides multi-modal sequencing methods that comprise performing a first template dependent sequencing reaction in a first mode comprising a first set of reaction conditions and collecting initial sequencing information produced by the first sequencing reaction. Additionally, the methods can include performing a second sequencing reaction of the template, or a copy thereof, in a second mode that includes a second

set of reaction conditions and collecting additional sequencing information produced by the second sequencing reaction. The methods can include compiling the initial and additional sequencing information to provide a sequence of at least a portion of the template.

[0019] The second sequencing reaction of the methods described above is optionally produced by altering one or more reaction conditions of the first sequencing reaction, and the initial and additional sequencing information are collected in real time. Optionally, altering one or more reaction conditions comprises adding one or more polymerase cofactors to the first sequencing reaction, where the cofactors of the first sequencing reaction are optionally Mn^{++} or Mg^{++}. Both the first and second sequencing reactions comprise single template molecule sequencing reactions. For example, the first mode can produce a higher branch fraction than the second mode. The second mode optionally produces longer read lengths than the first mode. The template of the methods described above is optionally adapted to sequencing, e.g., a single circular template molecule, e.g., where the method includes switching between the first and second modes.

[0020] The invention also provides methods for determining which of two or more labeled nucleotides is incorporated at a site during a template-dependent polymerization reaction. The methods include incorporating the nucleotide into a nucleic acid polymer produced by the polymerization reaction, whereby signal pulses or signatures are generated. The methods further include 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. The multiple pulses optionally include 2 to 20 pulses and can be generated from branch fraction nonincorporation events, which events are optionally induced by sequencing compositions that include at least one species of metal ion. Metal ions of the method can include Mg^{++}, Mn^{++}, Zn^{++}, Co^{++}, Ca^{++}, Fe^{++}, Cr^{++} and Sr^{++}. Optionally, using the presence of multiple pulses comprises distinguishing incorporation and nonincorporation signals to assign which labeled nucleotide was incorporated at the site.

[0021] In another aspect, the invention provides nucleic acid sequencing systems that, during operation of the system, sequences a nucleic acid. The nucleic acid sequencing system comprises a signal detector that detects at least signal pulses or signatures from branch fraction nonincorporation events during sequencing of a template nucleic acid, system instructions or

software that assigns 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.

[0022] Signal pulses or signatures from branch fraction nonincorporation events of the nucleic acid sequencing systems described above are optionally generated during a first sequencing mode, where the signal detector subsequently detects signal pulses generated during a low branch fraction second sequencing mode, and where the system instructions assign a sequence based upon detection of signal pulses or signatures from the first and second sequencing modes.

[0023] The nucleic acid sequencing systems optionally comprise a zero-mode waveguide or nanohole proximal to the signal detector, where during operation of the system, a sequencing reaction is contained by the zero-mode waveguide or nanohole.

[0024] In another aspect, the invention provides compositions that include a modified recombinant nucleic acid polymerase that exhibits an altered property selected from an increased branching fraction during a polymerization by the polymerase, an altered translocation property of the polymerase during a polymerization reaction, and a combination of these two altered properties, where the altered property or properties is altered as compared to a corresponding wild-type polymerase.

[0025] The modified recombinant polymerase of the compositions described above can optionally be 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. Other available polymerases can also be used as starting points for modification to alter translocation rates or to modulate branch fraction activity, such as reverse transcriptases and DNA-dependent RNA polymerases.

[0026] The modified recombinant polymerase exhibiting an increased branching fraction can optionally comprise 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; where numbering of the residues is relative to a wild-type Φ29 polymerase of SEQ ID NO: 3.

The modified recombinant polymerase exhibiting an increased branching fraction can optionally exhibit increased exonuclease activity, where the polymerase exhibits an exonuclease rate that is about 10% to 50% as compared to its polymerization rate.

[0027] Modified recombinant polymerases of the compositions described above optionally exhibit a branching fraction that is at least 50% greater, at least 100% greater, or at least 200% greater than the branching fraction of a wild-type Φ29 polymerase of SEQ ID NO: 3. Optionally, the polymerases exhibit increased exonuclease activity as compared to the corresponding wild type polymerase, where the increased exonuclease activity is optionally about 10% to 50% as compared to its polymerization rate.

[0028] The modified recombinant polymerase exhibiting an altered translocation property can optionally comprise a fusion protein that comprises at least a subsequence of the parental polymerase (e.g., a Φ29 DNA polymerase) and at least one heterologous polypeptide sequence (see, e.g., SEQ ID No. 1 and SEQ ID No. 2 in the sequence listing herein). Optionally, the fusion of the at least a subsequence of the parental polymerase and the heterologous polypeptide sequence can occur at or near the c-terminal end of the parental polymerase. The wild-type polymerase is optionally a Φ29 polymerase. The heterologous polypeptide sequence can optionally comprise at least one charged amino acid, where the at least one charged amino acid can optionally be histidine or a chain of histidines. Optionally, the fusion proteins described above can comprise a linker between the at least a subsequence of the parental polymerase and the heterologous polypeptide sequence, where the linker optionally comprises a Ser3Gly linker.

[0029] The modified recombinant polymerase exhibiting an altered translocation property can optionally comprise 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; where numbering of the residue positions is relative to a wild-type Φ29 polymerase of SEQ ID NO: 3.

[0030] The altered translocation property of the modified recombinant DNA polymerases can optionally comprise a delay in translocation. Modified recombinant polymerases of the compositions described above can optionally exhibit a delay in translocation that is at least about 2.5x, 10x or 15x greater than a corresponding wild-type polymerase.

[0031] The modified recombinant polymerases of the compositions described above optionally exhibit an increased nucleotide or nucleotide analog residence time or increased processivity as compared to a corresponding wild-type polymerase.

[0032] The compositions comprising a modified recombinant polymerase that exhibits an altered property described above can include a phosphate-labeled nucleotide analog, a DNA template, and a modified recombinant DNA polymerase, e.g., any of the polymerases described above, that can incorporate the nucleotide analog into a copy nucleic acid in response to the DNA template. These compositions can be present in a DNA sequencing system, e.g., a zero-mode waveguide or nanohole. Optionally, the polymerase of the compositions can be immobilized on a surface.

[0033] In a related aspect, the invention provides methods of sequencing a nucleic acid template. The methods include providing a reaction mixture that includes the nucleic acid template, a replication initiating moiety that complexes with or is integral to the template, the modified recombinant nucleic acid polymerase of the compositions described above, where the polymerase is capable of replicating at least a portion of the template using the moiety in a template-dependent polymerization reaction, and one or more nucleotides and/or nucleotide analogs. In addition, the methods subject 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, where one or more nucleotides and/or nucleotide analogs are incorporated into the resulting copy nucleic acid. The methods additionally identify a time sequence of incorporation of the one or more nucleotides and/or nucleotide analogs into the resulting copy nucleic acid. Optionally, the methods include a modified recombinant polymerase that exhibits increased processivity relative to the wild-type polymerase. The methods optionally include identifying the time sequence of incorporation by observing more than one signal pulse per nucleotide incorporation event. Subjecting the reaction mixture to a

polymerization reaction and identifying a time of sequence incorporation can optionally be performed in a zero mode waveguide, nanohole or other micro- or nano-structure.

[0034] The invention also provides methods of making a nucleic acid that include providing a reaction mixture that comprises a template, a replication initiating moiety that complexes with or is integral to the template, a modified recombinant DNA polymerase with an altered property or combination of altered properties, e.g., such as those described above, which can replicate at least a portion of the template using the moiety in a template-dependent polymerase reaction, and one or more nucleotides and/or nucleotide analogs. In addition, the methods include 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. Optionally, the methods include detecting incorporation of at least one of the nucleotides and/or nucleotide analogs, which optionally includes observing more than one signal pulse per nucleotide incorporation event. The mixture is optionally reacted in a zero mode waveguide or nanohole, and the modified recombinant polymerase optionally exhibits an increased nucleotide or nucleotide analog residence time and/or processivity as compared to the parental polymerase.

[0035] In a related aspect, the invention provides methods of making a modified recombinant DNA polymerase that include mutating a polymerase of interest, e.g., a Φ29-type DNA polymerase, and selecting resulting modified polymerases for a property selected from increased branching fraction and altered translocation. Mutating the polymerase of interest can optionally comprise structurally modeling the polymerase to identify a feature that may affect branch fraction or altered translocation. Optionally, mutating the polymerase of interest includes making a library of modified recombinant polymerases, and selecting the modified polymerases includes screening the library to identify at least one member exhibiting the property. The polymerase of interest optionally includes 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. Modified recombinant polymerases that exhibit the property optionally exhibit increased nucleotide or nucleotide analog residence time, or increased processivity, as compared to a corresponding wild-type polymerase.

BRIEF DESCRIPTION OF THE DRAWINGS

Figure 1 schematically illustrates a system of the invention.

Figure 2 schematically illustrates a sequencing by incorporation reaction and the resulting characteristics of signal pulses detected by a system that employs reaction conditions or polymerases of the invention.

Figure 3 is a time sequence of signal pulses generated from a sequencing by incorporation reaction under relatively high branch fraction conditions.

Figure 4 is a schematic illustration of a sequencing by incorporation reaction in which unincorporatable nucleotides are included in the reaction.

Figure 5 shows a theoretical representation of the frequency of binding events per incorporation for a polymerase reaction having one rate-limiting step or two rate-limiting steps within an observable phase.

Figure 6 is a schematic illustration of the reaction cycle for polymerase-mediated nucleic acid primer extension.

Figure 7 schematically illustrates a system of the invention that utilizes more than one mode of sequencing.

DETAILED DESCRIPTION

[0036] The invention is generally directed to modified or engineered compositions that are characterized by modified profiles or characteristics for incorporation of nucleotides in template directed nucleic acid synthesis. Such characteristics include, for example, increased frequency of branching events, changes in reaction rates that lead, e.g., to delayed polymerase translocation and/or increased nucleotide or nucleotide analog retention time during polymerization reactions. Individually or in combination, these modifications can increase sequence readout accuracy (e.g., increase sequence accuracy in single molecule sequencing reactions) using the methods of the invention. Polymerases of the invention optionally also include additional mutations or modifications that provide other desirable features, e.g., modify one or more kinetic features of the polymerase (e.g., increased processivity), increased surface stability for polymerases bound to a surface, or the like.

[0037] During sequencing by incorporation, e.g., single molecule sequencing by synthesis (SMS), nucleotide (or nucleotide analog) incorporation events are detected in real-time as the bases are incorporated into the extension product. This can be accomplished by immobilizing a synthesis complex, which includes a polymerase enzyme, such as a DNA polymerase enzyme, a template nucleic acid sequence, and a primer sequence that is complementary to a portion of the template sequence, within an optically confined space or otherwise resolved as an individual molecular complex. Some SMS methods employ nucleotide analogs that include fluorescent labels coupled to the polyphosphate chain of the analog, which are then exposed to the complex. Upon incorporation, the nucleotide – along with its fluorescent label – is retained by the complex for a time and in a manner that permits the detection by a sequencing system of a signal "pulse" from the fluorescent label at the incorporation site. The sequentially detected signal pulses are then interpreted by the sequencing system to generate a readout corresponding to the sequence of the template nucleic acid. For a discussion of preferred sequence by incorporation processes, see, e.g., U.S. Patent Nos. 6,056,661, 7,052,847, 7,033,764, 7,056,676, 7,361,466, the full disclosures of which are hereby incorporated herein by reference in their entirety for all purposes. Further details regarding base calling during sequencing by incorporation methods are found in Tomaney et al. PCT Application Serial No. PCT/US2008/065996 METHODS AND PROCESSES FOR CALLING BASES IN SEQUENCING BY INCORPORATION METHODS, incorporated herein by reference in its entirety for all purposes.

[0038] **Figure 2** schematically illustrates a sequencing by incorporation reaction and the resulting patterns of signal pulses detected by a system that employs reaction conditions or polymerases of the invention. Figure 2A schematically illustrates a polymerization reaction where dye-labeled nucleotides are incorporated in a stepwise fashion according to the sequence of the template strand. When a dye-labeled nucleotide enters the detection region (dashed box) which encompasses the polymerase, the dye emits optical signal pulses or signatures in response to excitation radiation that are detected by a signal detector. Examples of detection methods and optically confined reaction regions include, e.g., Zero Mode Waveguides, e.g., as described in U.S. Patent Nos. 6,917,726, 7,013,054, 7,181,122, and 7,292,742, the full disclosures of which are hereby incorporated by reference in their entirety for all purposes.

Figure 2B schematically illustrates the patterns or characteristics of signal pulses that would arise under standard conditions (Panel I), increased branching conditions (Panel II), conditions that include a polymerase with a decreased translocation rate (Panel III) and conditions that include a polymerase that exhibits increased nucleotide analog residence time (Panel IV). The resulting patterns or characteristics of signal pulses from the various conditions are described in detail below.

I. INCREASED BRANCHING

[0039] "Branching" is a phenomenon that occurs during polymerization. During a polymerase kinetic cycle, sampling of each of four possible nucleotides (or nucleotide analogs) occurs until a correct Watson-Crick pairing is generated (see, e.g., Hanzel et al. WO 2007/076057 POLYMERASES FOR NUCLEOTIDE ANALOG INCORPORATION for an example model description of the kinetic cycle of a polymerase). However, chemical linkages between a sampled nucleotide and a 3'OH group of a preceding base can fail to occur for a correctly paired nucleotide, due, e.g., to release of the correctly paired base from the active site. This can occur as a result of the nucleotide leaving the site without a covalent bond being formed, or e.g., as a result of cleavage of the covalent bond (e.g., due to exonuclease activity) prior to polymerase translocation to the next incorporation site. During single molecule sequencing (SMS) procedures, and particularly those single molecule processes that monitor incorporation in real time, where both the failed incorporation and the actual incorporation of the nucleotides provide signal pulses, sequences deciphered for the incorporation site can have an incorrect "insertion" relative to the correct sequence as a result of such branching. This phenomenon is termed "branching" because it leads to a branch in the sequence (a site where two identical molecules will be read as having different sequences) and may lead to increased error rates during SMS.

[0040] While branching can, in many applications of single molecule sequencing processes, be viewed as an accuracy reducing phenomenon, in at least a first aspect of the present invention, increased branching is exploited to increase sequence accuracy by providing redundant signal events resulting from iterative sampling of labeled nucleotides or

nucleotide analogs. In particular, improved sequence reliability and accuracy is achieved by providing reaction conditions and/or polymerases that exhibit a relatively high average branching fraction for a particular nucleotide or nucleotide analog and a certain distribution of branch signal pulses around this average. Such compositions are used in combination with a sequencing system that observes and interprets more than one signal pulse or signature per incorporation event to identify the nucleotide sequence of a target or template nucleic acid. This is advantageous in the present invention because detecting more than one signal pulse or signature per incorporation event provides inherent redundancy of signal for each desired incorporation event. In some cases, the "signature" will include regions of optical signal versus time that are characteristic of the branching nucleotide, but do not appear as individual pulses. This can occur, for example, when a sequence of pulses are not individually resolved. Further details regarding sequencing under high branch fraction conditions can be found in Bjornson et al. PCT Application Serial No. PCT/US2009/000921 COMPOSITIONS AND METHODS FOR USE IN ANALYTICAL REACTIONS, incorporated herein by reference in its entirety for all purposes. Additional information useful to sequencing under high branch fraction conditions can be found in Bjornson et al. PCT Application Serial Number PCT/US2009/002003 TWO SLOW-STEP POLYMERASE ENZYME SYSTEMS AND METHODS, incorporated herein by reference in its entirety for all purposes.

[0041] The branching fraction is the proportion of cognate nucleotide (or nucleotide analog, e.g., A488dA4P) dissociation events from the polymerase active site as compared to the total number of events, e.g., the sum of the dissociation events and the incorporation events for the cognate nucleotide or nucleotide analog. The present invention provides high branch fraction polymerization reactions. As used herein, a high branch fraction polymerization reaction includes a reaction that exhibits a branching fraction of at least about 70% or more, about 80% or more, about 85% or more, about 90% or more, or about 95% or more. For example, in a sequencing reaction in which the branching fraction is 80%, 80% of the total interactions of the nucleotide or nucleotide analog with the polymerase binding pocket result in dissociation, rather than incorporation, of the nucleotide or nucleotide analog.

[0042] An aspect of the invention is a method of nucleic acid sequencing by monitoring an optical signal from a polymerase reaction, wherein the base call, or the assignment of the incorporated base is made on the basis of multiple pulses from the same nucleotide. The number of pulses used to assign which base has been incorporated may depend on the branching fraction under the conditions of the polymerase reaction. In some cases, the number of pulses used to assign which nucleotide has been incorporated will vary between the different nucleotides in that reaction medium. The number of pulses used to assign which nucleotide is incorporated can be about 2, 3, 4, 5, 6, 7, 8, 9, 10, 12, 15, 20, 30, or more pulses. The number of pulses can be between about 2 and about 30 pulses, between about 2 and about 20 pulses, or between about 3 and about 25 pulses. The number of pulses can be expressed, for example, as the average number of pulses used to assign a given nucleotide for one sequencing reaction.

[0043] Sequence read errors during SMS can also derive from the incorporation of nucleotides or nucleotide analogs that constitute dark matter (for the purposes of this disclosure, "dark matter" refers to unlabeled nucleotides or nucleotide analogs with nonfunctional labels, e.g., broken fluorophores). Here, a genuine incorporation event is not detected due to the absence of a signal pulse from the dark matter, and a subsequent incorporation event is interpreted by the sequencing system as occurring at the position where the dark matter was incorporated. Dark matter, therefore, may potentially contribute to error rates in single molecule sequencing that utilizes the incorporation of labeled nucleotides.

[0044] In certain aspects, the reaction conditions, modified recombinant polymerases, and/or nucleotide analogs of the present invention – employed in conjunction with the sequencing system of the present invention – reduce sequence read errors that might result from missed pulses. The reaction conditions induce a relatively high branch fraction polymerization reaction – and the modified recombinant polymerases exhibit increased average branching fractions - such that a greater number of nucleotide analogs, which, if incorporated, would correctly pair with the corresponding nucleotide of the template strand, enter the active site before an analog is eventually incorporated into the extension product. The nucleotide analogs that enter the active site, but fail to incorporate, produce redundant signal pulses or signatures at each incorporation site, resulting in multiple redundant signal

events for each incorporation event. An example of signal pulses generated under relatively high branch fraction conditions is shown in **Figure 3.**

[0045] The sequencing system takes into account the average branching fraction of the polymerase and a certain distribution of branch pulses or signatures per nucleotide incorporation around this average. Because multiple signal pulses are observed for each incorporation event, branching events involving unlabeled nucleotides or nucleotides with nonfunctional labels, i.e., dark matter, do not result in a sequencing read error, but rather only slightly decrease the distribution of the average number of pulses or signatures per incorporation. In the event that dark matter is incorporated into the extension product, signal pulses derived from branching events involving nucleotide analogs with functional labels prior to dark matter incorporation can provide sufficient redundancy for determining the correct base at the incorporation site.

[0046] As will be appreciated, high branch fraction sequencing conditions can also be used for sequencing RNA templates, for example using reverse transcriptase enzymes and for RNA synthesis, for example by DNA dependent RNA polymerases.

A. Enhanced Sequencing Using Reaction Conditions That Promote Branching

[0047] The present invention provides reaction conditions – such as the type, level, and relative amounts of cofactors – that increase the frequency of branching events during nucleic acid polymerization reactions. Such reaction conditions may be used in combination with polymerases that are engineered to exhibit increased branching fractions under selected conditions, or can be used with polymerases that are unaltered with respect to branching properties. The phosphoryl transfer reaction of DNA polymerases is typically catalyzed by a two-metal ion mechanism, where two divalent metal ions, e.g., Mg^{++} and/or Mn^{++}, complexed with the DNA polymerase facilitate the incorporation of a nucleotide into the 3'OH of the extension product. One of the metal ions is proposed to interact with the 3'OH of the primer strand, thereby facilitating its attack on the α-phosphate of the incoming nucleotide. Both metal ions are believed to stabilize the transition state that occurs during the course of the extension reaction.

[0048] During the course of the polymerase reaction, divalent metal cofactors, such as magnesium or manganese, will interact with the enzyme-substrate complex, playing a structural role in the definition of the active site. For a discussion of metal cofactor interaction in polymerase reactions, see, e.g., Arndt, et al., Biochemistry (2001) 40:5368-5375. For example, and without being bound to any particular theory of operation, it is understood that metal cofactor binding in and around the active site serves to stabilize binding of incoming nucleotides. For further details regarding the effect of metal cofactors on polymerase kinetics and nucleic acid synthesis reactions, see Bjornson et al. PCT Application Serial Number PCT/US2009/002003 TWO SLOW-STEP POLYMERASE ENZYME SYSTEMS AND METHODS, incorporated herein by reference in its entirety for all purposes.

[0049] In the context of the present invention, it has been discovered that modulation of the concentration of a divalent metal cofactor, or competitive modulation of two or more divalent metal cofactors, to the synthesis reaction can result in increased branching for enhanced nucleic acid sequencing without a consequent increase in negative reaction events. As described in detail herein, the increased branching provides redundant signal pulses or signatures, thereby reducing or eliminating the occurrence of missed signal pulses and improving sequence accuracy. As used herein, a signature can include regions of optical signal versus time that is characteristic of the branching nucleotide, but does not appear as an individual signal pulse.

[0050] In the synthesis reaction, certain divalent or trivalent metal cofactors, such as magnesium and manganese are known to interact with the polymerase to modulate the progress of the reaction (See, e.g., U.S. Patent No. 5,409,811). As will be appreciated, depending upon the nature of the polymerization reaction, environmental conditions (e.g., temperature, pH, etc.), the polymerase used, the nucleotides employed, etc., different metal co-factors will have widely varying catalytic effects upon the polymerization reaction. In the context of the present invention, different metal cofactors will be referred to herein based upon their relative catalytic impact on the polymerization reaction, as compared to a different metal included under the same reaction conditions. For purposes of discussion, a first metal cofactor that interacts with the polymerase complex to support the polymerization reaction to a higher level than a second metal cofactor under the same conditions is termed a "catalytic metal ion" or "catalytic metal."

[0051] The present invention provides sequencing compositions and methods that include, e.g., divalent metal ions at concentrations that induce high branch fraction polymerization reactions. Divalent metal ions of the invention can be, e.g., Mg^{++}, Mn^{++}, Zn^{++}, Co^{++}, Ca^{++}, Fe^{++}, Cr^{++}, and/or Sr^{++}. For the purposes of this disclosure, a high branch fraction polymerization reaction includes a reaction that exhibits a branching fraction of about 70% or more, about 80% or more, about 85% or more, about 90% or more, or about 95% or more.

[0052] In one embodiment, a high branch fraction polymerization reaction is induced in a sequencing composition that includes a DNA polymerase, e.g., a Φ29-derived polymerase that uses Mn^{++} as the sole source of metal cofactor at a concentration of about 250 μM or less, about 200 μM or less, about 150 μM or less, about 125 μM or less, about 100 μM or less, about 75 μM or less, or about 50 μM or less. By contrast, branching is typically not promoted in sequencing compositions that include, e.g., about 500 μM Mn^{++} or more, in the absence of other factors.

[0053] In another embodiment, a high branch fraction polymerization reaction is induced by a sequencing composition that includes a DNA polymerase, e.g., a Φ29-derived polymerase that uses Mg^{++} as the sole of metal cofactor at a concentration of about 1 mM or more, about 2 mM or more, about 3 mM or more, about 5 mM or more, about 10 mM or more, about 20 mM or more, about 30 mM or more, about 40 mM or more, or about 50 mM or more.

[0054] In yet another embodiment, a high branch fraction polymerization reaction is induced by a sequencing composition that includes a DNA polymerase, e.g., a Φ29-derived polymerase that uses both Mg^{++} and Mn^{++} as metal cofactors, where both the absolute values and ratio of the two concentrations of Mg^{++} and Mn^{++} determines the extent of branching. For example, the present invention provides sequencing compositions that induce branching by including both Mg^{++} and Mn^{++}, where the concentration of Mg^{++} is greater than the concentration of Mn^{++}. In one particular embodiment, the sequencing composition includes about 10 mM $MgCl_2$ and about 100 μM $MnCl_2$. A range of suitable concentrations to increase branching are 0-200 mM $MgCl_2$ and 0.01-50 mM $MnCl_2$, and all possible combinations of values between those two ranges.

[0055] In another embodiment, the reaction conditions include Mn^{++} and a metal cofactor other than Mg^{++}, such as calcium, barium, strontium, iron, cobalt, nickel, tin, zinc, and europium. For example, these metals can be added to the polymerization reaction in salt form such as $Sr(OAc)_2$, $CoCl_2$, $SnCl_2$, $CaCl_2$, or $ZnSO_4$. Both the absolute values and ratio of the two concentrations can influence the extent of branching.

[0056] The present invention also provides methods for increasing the branching fraction during sequencing by incorporation by modifying reaction conditions other than the identity and/or concentrations of metal ions. For example, the pH (lowering the pH to about 6.5), temperature (e.g., decreased temperature), addition of D_2O, and/or addition of small molecule inhibitors (e.g., a noncompetitive inhibitor that slows covalent attachment of the nucleotide to the 3-OH of the growing strand of the nucleic acid being synthesized, e.g., a non-competitive HIV-RT inhibitor), can be used to alter the branching fraction of the polymerization reaction.

[0057] As will be appreciated, the sequencing compositions and methods described above that utilize metal cofactors to induce relatively high branching fractions can be used in combination with any other embodiments described herein, including: (1) modified recombinant polymerases that exhibit increased branching fractions as compared to the corresponding wild-type polymerases; (2) iterative sampling of unincorporatable nucleotides; (3) two slow-step enzyme systems; (4) detection of noncognate branching events; (5) modified recombinant polymerases that exhibit altered translocation properties as compared to the corresponding wild-type polymerases; (6) modified recombinant polymerases that exhibit increased nucleotide residence time; (7) sequencing of nucleic acid templates using more than one mode; and/or (8) any other combination of embodiments described herein. As will also be appreciated, high branch fraction sequencing conditions can be used for sequencing RNA templates, for example using reverse transcriptase enzymes, and for RNA synthesis, for example by DNA dependent RNA polymerases.

B. Enhanced Sequencing Using Modified Recombinant Polymerases with Increased Branching Fractions

[0058] During a polymerase kinetic cycle, sampling of each of the possible nucleotides or nucleotide analogs occurs until a correct Watson-Crick pairing is generated. According to structural studies of DNA polymerases complexed with DNA substrates, the primer-terminus does not typically form a covalent bond with an incorrectly paired nucleotide (Berman, et al. (2007) "Structures of phi29 polymerase complexed with substrate: the mechanism of translocation in polymerases." *EMBO J* **26:** 3494–3505). Conversely, branching events can occur during the polymerase kinetic cycle, where chemical linkages between a correctly paired nucleotide and a 3'OH of a preceding base fail to form, e.g., due to premature release of the sampled nucleotide from the active site. The kinetic cycle is then repeated for the same site, eventually resulting in the physical incorporation of the correct nucleotide.

[0059] As described above, these branching events can result in sequence read errors in standard sequencing methods, e.g., due to extra incorporation signals relative to the template sequence, received by a sequencing system that monitors signal pulses from the nucleotide analog at the active site as a proxy for incorporation, if the system does not account for the branching events. However, a sequencing system that utilizes branching events and calls bases according to multiple signal pulses or signatures for each incorporation event can be used in combination with polymerases that exhibit a high average branch fraction to improve sequence read accuracy. Under such conditions, redundant signals pulses or signatures resulting from iterative sampling of a labeled cognate nucleotide or nucleotide analog can reduce the error rate of a sequence read, as compared to lower branch fraction sequencing conditions where only one or a small number of signal pulses for each incorporation event can go undetected by the sequencing system.

[0060] The present invention provides modified recombinant polymerases and reaction conditions with increased branching fractions that can be used to improve sequence read accuracy. The branching fraction is the proportion of cognate nucleotide (or nucleotide analog, e.g., A488dA4P) dissociation events from the polymerase active site as compared to the total number of events, e.g., the sum of the incorporation events and dissociation events for the cognate nucleotide or nucleotide analog. Either incorporation or non-incorporation events, or both, can be detected by monitoring a signal profile produced by a sequencing reaction.

[0061] In the present invention, modification of a DNA polymerase by mutagenesis is used to increase the frequency of branching events. In exemplary embodiments, this modification may include one or more of either creating a more loosely structured binding pocket for the (typically non-natural) nucleotides that are incorporated during SMS, or by structurally modifying the polymerase to increase exonuclease activity. Random mutation strategies can also be used, e.g., in conjunction with appropriate screening steps to select libraries of mutants for increased branching (or other properties of interest). Combinations of random and site-directed mutagenesis can also be used, typically in conjunction with selection of mutant libraries for a property of interest.

[0062] As will be appreciated, polymerase enzymes of the present invention are not limited to DNA polymerases. The present invention provides modified recombinant reverse transcriptase enzymes and modified recombinant DNA-dependent RNA polymerases, which can exhibit increased branching fractions during RNA template sequencing and RNA synthesis, respectively.

[0063] One class of example mutants described in this application were designed to address branching fraction by modifying various sites in, e.g., a Φ29 polymerase, predominantly in and around the binding pocket, to create weaker polymerase-analog interactions during an extension (polymerization) reaction. A second class of example mutants described in this application were designed to increase branching by modifying various sites in, e.g., a Φ29 polymerase, predominantly in and around the exonuclease domain in order to increase the exonuclease rate of the polymerase to about 10% to 50% as compared to its polymerization rate. As noted, the "branching fraction" is the proportion of cognate nucleotide (or nucleotide analog, e.g., a dye-labeled analog) dissociation events from the polymerase active site to the total number of events, e.g., the sum of the incorporation events and dissociation events. For the purposes of this disclosure, dissociation events also include cleavage of an incorporated nucleotide as a result of exonuclease activity. These mutational features, i.e., increased branching by creating a more loosely structured binding pocket or increasing exonuclease activity, can be provided in combination.

[0064] Desirably, the branching fraction for a polymerase for a given nucleotide of interest (e.g., a labeled nucleotide analog) can be more than 50%, more preferably more than 60%, yet more preferably more than 70%, and still more preferably more than 80% or more of the total interactions, e.g., dissociation events and association events, of the nucleotide analog with the polymerase binding pocket. In comparison, a parental Φ29 polymerase exhibits a branching fraction of approximately 23% for, e.g., a thymine nucleotide analog that includes an Alexa568 fluorophore (Invitrogen Inc., Carlsbad, CA) linked to the terminal phosphate of a hexaphosphate chain, also referred to as A568dT6P, wherein approximately 23% of the total events with a gamma-linked A568dT6P nucleotide analog in the polymerase binding pocket are dissociation events.

[0065] The invention provides methods for generating recombinant polymerases that comprise modifications that increase the frequency of branching, which can be useful in any number of applications where accuracy of polymerization is beneficial, e.g., high-throughput sequencing systems, e.g., in a nanohole (an aperture of less than 1 μM diameter through which a synthesis complex can be illuminated by optical energy or monitored electrochemically) or specialized nanoholes such as zero-mode waveguides (ZMW), SNP genotyping using single base extension methods, real time monitoring of amplification, e.g., RT-PCR methods, and the like. Also provided by the invention are compositions that include such polymerases and methods in which these polymerases can be useful in, e.g., sequencing or making DNA.

[0066] In some embodiments, the compositions can also include nucleotide analogs, and preferably, optically labeled, e.g., fluorescently labeled, nucleotide analogs. In particularly preferred aspects, the compositions will include one or more types of phosphate-labeled nucleotide analog or analogs, e.g., a nucleotide analog comprising from 3-7 phosphate groups which in preferred cases may include a fluorophore coupled to the phosphate chain that is released upon incorporation, which can be incorporated into a copy nucleic acid by the modified polymerase in response to a DNA template. In some embodiments, the compositions can be present in a sequencing system, e.g. in a nanohole or specialized nanohole such as a zero-mode waveguide, where a polymerase of the invention can optionally be immobilized on a surface.

[0067] Modification of a polymerase, e.g., any of the polymerases described herein, or polymerases homologous to those described herein, by any one or more the strategies described herein can increase the frequency of branching events by creating a more loosely structured binding pocket for non-native nucleotides. The modified polymerases can comprise at least one amino acid substitution or a combination of amino acid substitutions relative to the parental polymerase.

[0068] Polymerases of the invention can be described or analyzed based upon comparison to a reference (e.g., parental) polymerase. For example, a parental polymerase (e.g., a wild type polymerase from which the polymerase of the invention is derived, or an available mutant) can serve as a reference polymerase. Comparisons between a reference polymerase and a polymerase of the invention are performed under selected reaction conditions. In general, standard reaction conditions can be defined based on the usual reaction conditions (e.g., optimized for the reaction at issue) for a given parental (e.g., wild-type) polymerase. That is, the reaction condition preferences for many polymerases are known; the appropriate adaptations of these conditions to specific applications such as SMS are known or can be determined through routine optimization or reaction conditions, and these "application optimized" reaction conditions can be used for comparison of the polymerase of the invention to the reference polymerase. For example, reaction conditions can be optimized for a reference polymerase such as a Φ29 polymerase, e.g., in an SMS application, with the reaction conditions being selected for optimal processivity, optimal fidelity, increased or decreased branch fraction, or a combination thereof, with that reference polymerase being used in a comparison to a polymerase of the invention under the optimized conditions. For example, in one SMS application, the reaction conditions can include those described above, and comparison to the polymerase of the invention can be conducted by performing a comparative assay, using the methods herein.

[0069] A number of specific examples of modified polymerases, e.g. modified to increase the average branching fraction, are described herein. The binding pocket is a portion of the polymerase that encompasses the nucleotide binding site and analog base during the pairing of a nucleotide analog with a template DNA. Because of the physical proximity of the binding pocket to the incoming nucleotide or nucleotide analog, mutations to this region can

affect the branching fraction. However, mutations that increase the branching fraction are not limited to this area of the polymerase. For example, relative to a parental or wild-type Φ29 DNA polymerase, useful modifications can include any of the following mutations: N62D and Y454A; D362S; Y259H; F237Y; L381I; Y369H; H461Y; A377G; K138Q; H461D; A377S; N62D and K371Q; V118L; and K124R. For the purposes of this disclosure, a parental polymerase includes a wild-type or available mutant/recombinant polymerase which is additionally altered to produce the desired properties of the invention, e.g., increased branching, delayed translocation or increased nucleotide analog residence time. A list of specific useful Φ29 mutants and their branching fractions (% BF) and corresponding increases in branching fraction (% increase BF) for a particular 6P nucleotide analog relative to a reference Φ29 polymerase (N62D) is provided in Table A below.

Table A

	A568dT6P		A647dG6P	
Mutation(s)	%BF	% Increase BF	%BF	% Increase BF
N62D	23	-	38	-
N62D_Y454A	30	30	48	25
D362S	31	34	46	20
Y259H	33	44	52	37
F237Y	35	53	46	21
L381I	37	60	47	22
Y369H	40	75	55	43
H461Y	43	86	57	49
A377G	44	90	75	97
K138Q	45	96	61	59
H461D	46	102	62	62
A377S	61	167	70	84
N62D_K371Q	62	172	94	145
V118L	65	181	74	92
K124R	73	216	84	119
	A555dC6P		A660dA6P	
Mutation(s)	%BF	% Increase BF	%BF	% Increase BF
N62D	27	-	22	-
N62D_Y454A	38	42	23	3
D362S	39	45	26	20
Y259H	39	44	25	11
F237Y	41	52	25	11
L381I	37	36	26	18
Y369H	43	59	28	27

	A555dC6P		A660dA6P	
Mutation(s)	%BF	% Increase BF	%BF	% Increase BF
H461Y	50	86	31	42
A377G	37	36	32	47
K138Q	48	78	27	21
H461D	51	91	31	39
A377S	72	168	56	152
N62D_K371Q	87	224	82	272
V118L	68	154	47	114
K124R	79	195	63	183

[0070] As noted, the branching fraction, e.g., % branching, is a relative measure of the number of times a correctly paired base, e.g., a Watson-Crick paired base, leaves the active site of the polymerase without forming a phosphodiester bond with the 3'OH of the primer-terminus relative to the total number of interactions that occur between a nucleotide (or nucleotide analog) and the binding pocket of the polymerase, e.g., the total number of opportunities the nucleotide or nucleotide analog has to correctly pair and incorporate. Additionally, for the purposes of this disclosure, branching refers to cleavage and dissociation from the polymerase active site of an incorporated nucleotide as a result of exonuclease activity. Branching is expressed as a percentage of the dissociation events vs. the total sum events, e.g., dissociation and association events. For example, for a polymerase harboring the Y369H mutation, for every 100 times an A568dT6P analog (i.e., a thymidine hexaphosphate nucleotide in which the terminal phosphate is labeled with an Alexa568 dye) interacts with the binding pocket of this polymerase, 40 of the events are non-productive dissociation events, e.g., wherein the analog dissociates from the polymerase instead of participating in a polymerization reaction. For this polymerase, the percent increase in branching fraction is 75% as compared to a reference phi29 polymerase (N62D) under identical reaction conditions.

[0071] The branching fraction can be measured by loading a polymerase active site with a cognate-matching nucleotide analog that can bind in the +1 and +2 positions. In the absence of divalent cation, this nucleotide cannot be incorporated into the DNA strand, so will pair with the template nucleotide at the +1 position but be released at some frequency specific for that analog/polymerase combination, e.g., the branching rate. This loading reaction is then followed by addition of a divalent cation that supports extension, e.g., Mn^{2+}, and a terminating–

type nucleotide analog, e.g., a dideoxynucleotide, comprising the same base as the cognate-matching analog in the loading step.

[0072] The dideoxy-analog will be incorporated into any +1 sites that are unoccupied and, once added, preclude further extension. Hence, polymerase active sites that are already occupied by a paired analog base extend to the +2 position, those that are not occupied (i.e. "branched") incorporate the dideoxy-type analog at +1 and do not extend, resulting in a single base addition. The extension products of this reaction are visualized by standard separation methods, e.g., gel or capillary electrophoresis, and the ratio of terminated product that is generated when a dideoxynucleotide is incorporated at the +1 position divided by the total terminated product, e.g., when a dideoxynucleotide is incorporated at both the +1 and +2 positions, indicates the fraction of 'branched' events that occur.

[0073] The branching fraction exhibited by modified polymerases of the present invention, e.g., a modified Φ29 polymerase, can be greater than a branching fraction exhibited by the corresponding wild-type polymerase for a given nucleotide analog. For example, a modified recombinant polymerase of the invention can exhibit an increased branching fraction that is greater than about 20% for the phosphate-labeled analog, greater than 50% for the phosphate-labeled analog, greater than 75% for the phosphate-labeled analog, greater than 100% for the phosphate-labeled analog, greater than 150% for the phosphate-labeled analog, or greater than 200% for the phosphate-labeled analog, as compared to the corresponding wild-type polymerase, e.g., a wild-type Φ29 polymerase, under the standard reaction conditions described above.

[0074] In some embodiments, the modified polymerase that exhibits an increased frequency of branching can also exhibit a K_m for a given phosphate-labeled nucleotide analog, e.g., any of the phosphate-labeled nucleotide analogs described herein, that is useful to achieve increased branching. For enzymes obeying simple Michaelis-Menten kinetics, kinetic parameters are readily derived from rates of catalysis measured at different substrate concentrations. The Michaelis-Menten equation, $V=V_{max}[S]([S]+K_m)^{-1}$ relates the concentration of uncombined substrate ([S], approximated by the total substrate concentration), the maximal rate (V_{max}, attained when the enzyme is saturated with substrate), and the

Michaelis constant (K_m, equal to the substrate concentration at which the reaction rate is half of its maximal value), to the reaction rate (V). To determine a K_m for a particular analog a series of extension reactions are performed with a varying concentration of the analog of interest with a fixed, saturating concentration of native nucleotides. A fit of the rate versus the substrate concentration generates estimation of the $-K_m$ as the slope of this line.

[0075] The present invention also provides polymerases with increased exonuclease activity for increasing the branching fraction during a template-dependent polymerization reaction, e.g., SMS. In a preferred aspect, a polymerases of the invention exhibits an exonuclease rate that is between about 10% and 50% as compared to its polymerization rate.

[0076] As will be appreciated that the above-identified modified or recombinant polymerases that display increased branching fractions may optionally include additional modifications that confer other useful properties described herein, e.g., delayed translocation, increased nucleotide analog residence time and/or increased processivity. As will also be appreciated, the above-identified modified or recombinant polymerases that display increased branching fractions can be used in combination with any other embodiments described herein, including: (1) reaction conditions that increase the frequency of branching fraction nonincorporation events; (2) iterative sampling of unincorporatable nucleotides; (3) two slow-step enzyme systems; (4) detection of noncognate branching events; (5) modified recombinant polymerases that exhibit altered translocation properties as compared to the corresponding wild-type polymerases; (6) modified recombinant polymerases that exhibit increased nucleotide residence time; (7) sequencing of nucleic acid templates using more than one mode; and/or (8) any other combination of embodiments described herein.

C. Enhanced Sequencing Using Iterative Sampling of Unincorporatable Nucleotide Analogs

[0077] The present invention also employs nucleotide based competitive reagent compositions for identifying sequence elements, despite not being incorporated in a nascent nucleic acid strand. In particular, the unincorporatable nucleotide analogs of the invention, while not being incorporatable, may be nonetheless capable of specifically associating with the polymerase enzyme. That is, the polymerase will sample the unincorporatable nucleotides,

retaining them within the active site for a greater length of time than nucleotides that are not complementary to the position in the template nucleic acid, and release them when they cannot be incorporated. By providing different types of nucleotide or nucleoside analogs, e.g., mimetics of A, G, T, C, and/or U, bearing distinguishable labels, e.g., spectrally resolvable fluorophores or other labeling groups, one can monitor the sampling of these nucleotides as an indication of the nucleotide that is next to be incorporated. For example, one may provide labeled, unincorporatable nucleotide analogs at concentrations in excess of incorporatable nucleotides, e.g., 2X, 5X or even 10X or greater. Each incorporation of an incorporatable nucleotide will, by virtue of the excess concentration, be preceded by repeated sampling events of the unincorporatable nucleotides, which will each carry its associated signal event. The incorporatable nucleotides may then either bear no label, or preferably, bear a label that is distinguishable from the unincorporatable nucleotides, so as to mark the termination of the sampling of a given base and proceeding onto the next base in the sequence. In such cases, it may be desirable to label all incorporatable nucleotides with a single type of fluorophore, i.e., indistinguishable from the label groups on the other types of incorporatable nucleotides present, but distinguishable from all of the unincorporatable nucleotides.

[0078] The signal detection for the foregoing process is schematically illustrated in **Figure 4**. In particular, Figure 4 shows a schematic illustration of a set of signal traces from a single molecule sequence by incorporation reaction. As shown, the plot shows five signal traces, one for each type of differentially labeled unincorporatable nucleotide analog (indicated as A', T', G' and C', as well as a trace for the signal associated with the type of label coupled to the incorporatable nucleotide (labeled as "I"). As shown, repeated sampling of the cognate unincorporatable nucleotide analog, e.g., A', provides an iterative set of signal events 402, followed by a signal 404 on the I trace indicating conclusion of the incorporation event. This pattern is repeated for the next base to be incorporated (indicated by iterative signals 406 in the T' trace, followed again by the incorporation signal 408, in the I trace, and again by the iterative sampling signal 410 in the A' trace followed by the incorporation signal 412 in the I trace. Because these unincorporatable nucleotides are mimetic of the base to be incorporated, they possess a longer retention time in the active site than the analog that is not complementary to the next base in the template, and as such, provide a signal profile that is distinguishable

from random, incorrect sampling, e.g., as indicated by transient signal events 414. Such iterative sampling may include two, three, four, five, ten or greater than ten signal events for each incorporation.

[0079] As noted above, the competitive reagents used are going to be non-reactive in the reaction of interest. In preferred aspects, and without being bound to any particular theory of operation, the competitive compounds may possess structures similar to nucleotides or portions thereof, such that they can competitively interact with the reaction of interest, e.g., through association with the polymerase active site. By way of example, such structures may comprise a polyphosphate component, e.g., a pyrophosphate, triphosphate, tetraphosphate, pentaphosphate, or longer phosphate chain, so that the compound mimics one or more of a nucleotide or the product of a polymerase mediated incorporation reaction, which is capable of competitively interacting with the polymerase, relative to the nucleotide analogs.

[0080] In certain preferred cases, additional components may be coupled to the polyphosphate component that mimic other portions of the nucleotide or nucleotide analog. By way of example, the polyphosphate component may be coupled to a cyclic and/or aromatic component that may structurally mimic the nucleoside component in its interaction with the polymerase. Such structures are generally illustrated by the following structure:

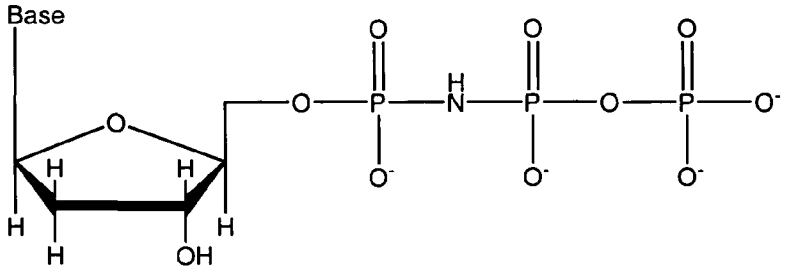
$$P\text{-}(P)_n\text{-}A;$$

where P is a phosphate or phosphonate group, n is an integer from 1 to about 6, and A includes a cycloalkyl or aryl group, a carbohydrate group, or the like.

[0081] In the case of nucleotide analogs used in analytical primer extension reactions, e.g., in nucleic acid sequence analysis, such nucleotide analogs will be unincorporatable in such primer extension reaction by the polymerase used. Further, in preferred aspects, such unincorporatable analogs will typically still be capable of interaction with the polymerase, e.g., active site binding, but will be unable to be incorporated in a primer extension reaction. In preferred aspects, this is accomplished by providing nucleotide analogs that possess nonhydrolyzable groups within the phosphate chain, such that the phosphoester linkage between the analog and the primer strand, cannot be formed, as mediated by the polymerase. One particularly effective approach to producing an unincorporatable nucleotide analog

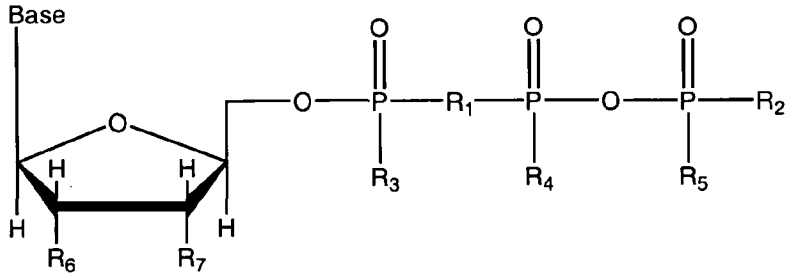
includes replacing the phosphoester linkage between the alpha and beta phosphate of a nucleoside polyphosphate with a nonhydrolyzable linkage.

[0082] One example of such an analog is illustrated below, where the oxygen group between the alpha and beta phosphate groups is replaced with an nonhydrolyzable linkage, such as the illustrated amino group.



[0083] Although illustrated as an amino linkage, it will be appreciated that a variety of other linkages may be used between the alpha and beta phosphates, e.g., an amino, methyl, thio, or other linkages not hydrolyzed by polymerase activity. Additionally, although illustrated as including three phosphate groups analogous to a nucleoside triphosphate, it will be appreciated that other polyphosphate configurations may be employed in the invention, including, for example, tetraphosphate analogs, pentaphosphate analogs, hexaphosphate analogs, and the like.

[0084] Thus, the structures employed in certain preferred aspects of the invention may generally be described with reference to the following structure:



where R_1 comprises a linking group that is non-hydrolyzable by the polymerase enzyme being used. Particularly preferred linkages include amino linkages, alkyl linkages, e.g., methyl, and thio linkages. While R_2 may comprise oxygen, in some preferred aspects, it will include additional phosphate groups, e.g., mono-, di-, or triphosphate groups coupled to the gamma

phosphate group. Alternatively or additionally, the R_2 group may include, in addition to or instead of additional phosphate groups, labeling functionalities that provide for the detection of the competitive substrates, but still permit its distinguishing from the incorporatable nucleotides. In other aspects, the R_2 group may include moieties that provide other functionalities to the reaction system other than as a labeling group. For example, R_2 may comprise an agent that reduces the potential for photodamaging effects on a polymerase enzyme, either coupled directly to the terminal phosphate group, or through a linking group.

[0085] The relative concentration of the competitive substrates to the incorporatable substrates, within a reaction mixture may generally be varied in accordance with a desired application. In particular, because the concentration of the competitive substrates affects the interactions of the complex with the incorporatable nucleotides, one can modulate those interactions by altering the ratios between incorporatable nucleotides and competitive substrates. In typical applications, however, the relative molar concentration of competitive substrate will range from about 0.5X to about 10X, 20X or greater of the concentration of the actual substrates (or incorporatable nucleotide analogs). Thus, the concentration ratio of unincorporatable nucleotide analogs to incorporatable nucleotide analogs will typically range from a lower ratio of from about 0.1:1, 0.2:1, 0.5:1 and 1:1, to an upper ratio of about 2:1, 3:1, 5:1, 10:1 or even 20:1, with each iteration of the foregoing being encompassed in the disclosure hereof.

[0086] As will be appreciated, iterative sampling of unincorporatable nucleotides can be used in combination with any other embodiments described herein, including: (1) reaction conditions that increase the frequency of branching fraction nonincorporation events; (2) modified recombinant polymerases that exhibit increased branching fractions as compared to the corresponding wild-type polymerases; (3) two slow-step enzyme systems; (4) detection of noncognate branching events; (5) modified recombinant polymerases that exhibit altered translocation properties as compared to the corresponding wild-type polymerases; (6) modified recombinant polymerases that exhibit increased nucleotide residence time; (7) sequencing of nucleic acid templates using more than one mode; and/or (8) any other combination of embodiments described herein.

D. Enhanced Sequencing Using Two Slow-Step Enzyme Systems Combined With Detection and Analysis of Branch Fraction Nonincorporation Events

[0087] Enzyme systems that exhibit kinetic mechanisms having two or more slow, kinetically observable, or partially rate-limiting reaction steps within an observable phase of the polymerase reaction can be useful for example, in single-molecule, real-time observations of such enzyme activity, which rely, at least in part, on detecting and identifying the enzyme reaction as it is occurring. By designing the reaction system to have two or more partially rate-limiting steps (i.e., "two slow-step" enzyme systems), the relative number of short, difficult to detect, events can be lowered. Details regarding enzyme systems exhibit kinetic mechanisms having two or more slow, kinetically observable, or partially rate-limiting reaction steps within an observable phase of the polymerase reaction can be found in Bjornson et al. PCT Application Serial Number PCT/US2009/002003 TWO SLOW-STEP POLYMERASE ENZYME SYSTEMS AND METHODS, incorporated herein by reference in its entirety for all purposes.

[0088] Certain types of template nucleic acid sequences present unique challenges during single molecule sequencing. For example, during single molecule sequencing of homonucleotide stretches (i.e., a portion of a template nucleic acid having two or more consecutive bases that are identical), ascertaining the number of nucleotide incorporation events that are represented by a series of signal pulses generated under high branch fraction sequencing conditions can be difficult using previous typical enzyme systems that exhibit one slow-step. By way of example, if 20 consecutive signal pulses are detected from a fluorescent-labeled nucleotide analog of dATP (e.g., A488dA4P, see below), a system that detects signal pulses and analyzes the time-sequence of those signal pulses for purposes of calling bases may not be able to accurately determine how many dATP analog incorporation events (i.e., the number of consecutive T residues in the template sequence) are represented by the 20 signal pulses. Accordingly, enzyme systems that permit a more predictable distribution of nucleotide binding events per incorporation are desirable, because, e.g., the number of nucleotides within a homonucleotide stretch can be determined by using a multiple of the expected number of binding events per incorporation.

[0089] In one aspect, the present invention provides two slow-step enzyme systems that provide a more predictable frequency or rate of nucleotide binding events (e.g., the number of times a nucleotide or nucleotide analog samples the active site of the polymerase) per incorporation event. While not being bound by any particular theory, the following theoretical basis is provided for obtaining improved single-molecule sequencing results, e.g., for homonucleotide stretches, by using a system having two or more slow steps within an observable phase. A model for the effect of two slow steps on the number of nucleotide binding events is described herein and illustrated in **Figure 5**. Figure 5 shows a plot of calculated number of binding events per incorporation for cases in which (1) one step is rate-limiting and (2) two equivalent partially rate-limiting (slow) steps are present for the observable phase in which the nucleotide is associated with the enzyme before unbinding.

[0090] For the case in which one step is rate-limiting, the distribution for the number of binding events per incorporation can be represented by the single exponential equation:

$$y = A_0 e^{-kt} \qquad \text{Eq. 1}$$

This represents the case in which, for example, one conformational change of the enzyme after nucleotide binding is the single slow step.

[0091] Figure 5 illustrates that where one slow-step is present in this phase, there is an exponentially decreasing number of binding events per incorporation, providing a distribution in which there is a relatively high probability that the number of binding events per incorporation will be low. In this scenario, it can be more challenging to distinguish homonucleotide stretches of the same base in the DNA template.

[0092] For the case in which there are two slow steps associated with binding of a nucleotide, for example via two consecutive conformational changes with similar rate constants, the number of binding events can be represented by a sum of two exponentials with an equation:

$$y = A_0 e^{-k_1 t} - B_0 e^{-k_2 t} \qquad \text{Eq. 2}$$

[0093] Figure 5 illustrates that for the case in which there are two slow steps, the probability of one or a low number of binding events per incorporation is relatively low as compared to enzyme systems having one slow step. In addition, the probability distribution for two slow steps exhibits a peak in the plot, with the most frequently observed number of binding events per incorporation greater than zero. This type of distribution can be advantageous for single-molecule sequencing where it is desired to resolve homonucleotide stretches. In this scenario, the expected number of binding events per incorporation will be a corresponding multiple factor of the most frequent occurrence of binding events per incorporation as governed by this distribution.

[0094] The two slow steps can include, e.g., nucleotide addition, enzymatic isomerization (such as to or from a closed state), and cofactor binding or release. The use of a distribution of pulses to determine a kinetic mechanism having two slow (kinetically observable) steps is described, for example, in Miyake et al. Analytical Chemistry 2008 80 (15), 6018-6022. The determination of the steps in a multistep reaction such as a polymerase reaction is described, for example, in Zhou, et al. J. Phys. Chem. B, 2007, 111, 13600-13610.

[0095] As noted above, the present invention provides enzyme systems that exhibit kinetic mechanisms having two or more slow, kinetically observable, or partially rate-limiting reaction steps within an observable phase of the polymerase reaction observable phase will generally have a time period during which the phase is observable. The time period for a bright phase, for example, can be represented by the pulse width. The time period for a dark phase can be represented, for example, by the interpulse distance. The length of each time period will not be the same for each nucleotide addition, resulting in a distribution of the length of the time periods. In some cases, the time periods with the shortest length will not be detected, leading to errors, for example in single-molecule sequencing. We have found that by designing enzyme systems such as polymerase reaction systems in which there are two slow, or kinetically observable, steps within an observable phase, the relative number of short, unobservable, time periods can be reduced, resulting in a higher proportion of observable sequencing events, and allowing for a more accurate determination of nucleotide sequence. As used herein, an observable phase includes phases that are not directly observable, but can be ascertained by measurements of other, related phases. For example, the lengths of dark phases can be

observed by measuring the times between optical pulses corresponding to a related bright optical phase. Also as described herein, a phase which is dark under some labeling conditions can be bright under other labeling conditions.

[0096] In natural polymerase-mediated nucleic acid synthesis, a complex is formed between a polymerase enzyme, a template nucleic acid sequence, and a priming sequence that serves as the point of initiation of the synthetic process. During synthesis, the polymerase samples nucleotide monomers from the reaction mix to determine their complementarity to the next base in the template sequence. When the sampled base is complementary to the next base, it is incorporated into the growing nascent strand. This process continues along the length of the template sequence to effectively duplicate that template. Although described in a simplified schematic fashion, the actual biochemical process of incorporation is relatively complex.

[0097] The process can be described as a sequence of steps, wherein each step can be characterized as having a particular forward and reverse reaction rate that can be represented by a rate constant. One representation of the incorporation biochemistry is provided in **Figure 6**. It is to be understood that the scheme shown in Figure 6 does not provide a unique representation of the process. In some cases, the process can be described using fewer steps. For example, the process is sometimes represented without inclusion of the enzyme isomerization steps 606 and 610. Alternatively, the process can be represented by including additional steps such as cofactor binding. Generally, steps which can be slow, and thus limit the rate of reaction will tend to be included. The present invention relates to methods, systems, and compositions in which the polymerization reaction has two or more slow steps within certain phases of the polymerase reaction. Various schemes can be used to represent a reaction having two slow steps that may have more or fewer identified steps. In some cases the two or more slow steps are consecutive. In some cases, there can be intervening fast steps between the two or more slow steps.

[0098] As shown in Figure 6, the synthesis process begins with the binding of the primed nucleic acid template (D) to the polymerase (P) at step 602. Nucleotide (N) binding with the complex occurs at step 604. Step 606 represents the isomerization of the polymerase from the open to closed configuration. Step 608 is the chemistry step where the nucleotide is

incorporated into the growing strand of the nucleic acid being synthesized. At step 610, polymerase isomerization occurs from the closed to the open position. The polyphosphate component that is cleaved upon incorporation is released from the complex at step 612. The polymerase then translocates on the template at step 614. As shown, the various steps can include reversible paths and may be characterized by the reaction constants shown in Figure 6 where:

[0099] k_{on}/k_{off}=DNA binding/release;

[0100] k_1/k_{-1}= nucleotide binding/release;

[0101] k_2/k_{-2}=polymerase isomerization (open/closed);

[0102] k_3/k_{-3}=nucleotide incorporation (chemistry);

[0103] k_4/k_{-4}=polymerase isomerization (closed/open);

[0104] k_5/k_{-5}=polyphosphate release/binding;

[0105] k_6/k_{-6}=polymerase translocation.

[0106] Thus, during steps 604 through 610, the nucleotide is retained within the overall complex, and during steps 604 and 606, reversal of the reaction step will yield an unproductive event, i.e., not resulting in incorporation. For example, a bound nucleotide at step 604 may be released regardless of whether it is the correct nucleotide for incorporation.

[0107] By selecting the appropriate polymerase enzyme, polymerase reaction conditions, and polymerase substrates, the absolute and relative rates of the various steps can be controlled. We have found that controlling the reaction such that the reaction exhibits two or more kinetically observable, or slow steps can produce a nucleic acid polymerization reaction in which the incorporation of the nucleotides can be observed more accurately. These characteristics are particularly useful for sequencing applications, and in particular single-molecule DNA sequencing.

[0108] In some cases, the invention involves a process having two or more kinetically observable steps that comprise steps after nucleotide binding through the step of product release. For the mechanism shown in Figure 6, this would be, for example, any of steps 606,

608, 610, and 612. In some cases, steps 608 (nucleotide incorporation) and 612 (product release) are the two slow, or kinetically observable steps. As noted previously, where one desires systems with slow steps in a dark phase, the invention may involve a process having two or more slow steps that comprise the steps after product release through nucleotide binding. For the mechanism shown in Figure 6, this would include steps 614 and 604.

[0109] In some cases, the invention involves a process in which there are two or more slow steps in two different observable phases within the polymerization, for example, two slow steps in a bright phase and two slow steps in a dark phase. For example, this could include a system having two slow steps in the steps after nucleotide binding through product release, and two slow steps for the steps after product release through nucleotide binding.

[0110] As is described herein, producing a process in which there are two or more slow steps in these portions of the polymerase reaction can result in a higher proportion of detectable enzyme states which can be useful, for example, to observe the sequential incorporation of nucleotides for nucleotide sequencing.

[0111] By the term slow-step we generally mean a kinetically observable step or partially rate-limiting step. The slow step need not be slow in the absolute sense, but will be relatively slow as compared with other steps in the enzymatic reaction. The slow, or kinetically observable steps, can be, for example, each partially rate-limiting, in that the rate of the step has a measurable effect on the kinetics of the enzymatic reaction. An enzymatic process, such as nucleic acid polymerization, can have both slower, kinetically observable steps and faster steps which can be so fast that they have no measurable effect on the kinetics, or rate, of the reaction. In some reactions, there can be a single rate-limiting step. For such reactions, the kinetics can be characterized by the rate of that single step. Other reactions will not have a single rate-limiting step, but will have two or more steps which are close enough in rate such that the characteristics of each will contribute to the kinetics of the reaction. A kinetically observable step is generally a step which is slow enough relative to the other steps in the reaction such that it can be experimentally ascertained. The experimental identification of a kinetically observable step can be done by the methods described herein, or by methods for assessing the kinetics of chemical and enzymatic reactions known in the art. For the current invention, the

slow, or kinetically observable steps, need not be the slowest step or the rate-limiting step of the reaction. For example, a process of the current invention can involve a reaction in which step 604, nucleotide addition is the slowest (rate-limiting) step, while two or more of steps 606, 608, 610, or 612 are each kinetically observable.

[0112] As used herein, the term rate, as applied to the steps of a reaction can refer to the average rate of reaction. For example, when observing a single-molecule reaction, there will generally be variations in the rates as each individual nucleotide is added to a growing nucleic acid. In such cases the rate of the reaction can be represented by observing a number of individual events, and combining the rates, for example, by obtaining an average of the rates.

[0113] As used herein, the reference to the rate of a step or rate constant for a step can refer to the forward reaction rate of the polymerase reaction. As is generally understood in the art, reaction steps can be characterized as having forward and reverse rate constants. For example, for step 608, k_3 represents the forward rate constant, and k_{-3} represents the reverse rate constant for the nucleotide incorporation. Some reaction steps, such as step 608, constitute steps which would be expected to be first order steps. Other steps, such as the forward reaction of step 604, with rate constant k_2, would be expected to be second order rate constants. For the purposes of the invention, for comparing the rate or the rate constant of a first order to a second order step, the second order rate constant k_2 can be treated as a pseudo-first order rate constant with the value [N]*k_2 where the concentration of nucleotide [N] is known.

[0114] It is generally desirable that the kinetically observable steps of the invention have rate constants that are lower than about 1000 per second. In some cases, the rate constants are lower than about 500 per second, lower than about 200 per second, lower than about 100 per second, lower than about 60 per second, lower than about 50 per second, lower than about 30 per second, lower than about 20 per second, lower than about 10 per second, lower than about 5 per second, lower than about 2 per second, or lower than about 1 per second.

[0115] In some embodiments the slowest of the two or more kinetically observable steps has a rate constant when measured under single-molecule conditions of between about 500 to about 0.1 per second, about 200 to about 0.1 per second, about 60 to about 0.5 per second, about 30 per second to about 2 per second, or about 10 to about 3 per second.

[0116] The ratio of the rate constants of each the two or more slow steps is generally greater than 1:10, in some cases the ratio of the rate constants is about 1:5, in some cases the ratio of the rate constants is about 1:2, in some cases, the ratio of rate constants is about 1:1. The ratio of the rate constants can be between about 1:10 and about 1:1, between about 1:5 and about 1:1, or between about 1:2 and about 1:1.

[0117] In some cases it is useful to consider the two slow-step system in terms of rates rather than rate constants. It is generally desirable that the kinetically observable steps of the invention have rates that are lower than about 1000 molecules per second when the reactions are carried out under single-molecule conditions. In some cases, the rates are lower than about 500 molecules per second, lower than about 200 molecules per second, lower than about 100 molecules per second, lower than about 60 molecules per second, lower than about 50 molecules per second, lower than about 30 molecules per second, lower than about 20 molecules per second, lower than about 10 molecules per second, lower than about 5 molecules per second, lower than about 2 molecules per second, or lower than about 1 molecule per second.

[0118] In some embodiments the slowest of the two or more kinetically observable steps has a rate when measured under single-molecule conditions of between about 500 to about 0.01 molecules per second, between about 200 to about 0.1 molecules per second, between about 60 to about 0.5 molecules per second, about 30 molecules per second to about 2 molecules per second, or about 10 to about 3 molecules per second.

[0119] The ratio of the rates of each the two or more slow steps is generally greater than 1:10, in some cases the ratio of the rates is about 1:5, in some cases the ratio of the rates is about 1:2, in some cases, the ratio of rates is about 1:1. The ratio can be between about 1:10 and about 1:1, between about 1:5 and about 1:1, or between about 1:2 and about 1:1.

[0120] A two or more slow-step system of the present invention can be obtained by selecting the correct set of polymerase enzyme, polymerase reaction conditions, and polymerase reaction substrates.

[0121] Table B presents exemplary Φ29 mutants that can exhibit two slow step behavior under appropriate reaction conditions. The first three modified polymerases exhibit

the most pronounced two slow step behavior, followed by the next six. As noted, the polymerases are optionally exonuclease-deficient; for example, they can also include an N62D substitution.

Table B

A484E/E375Y/K512Y/T368F
A484Y/E375Y/K512Y/T368F
N387L/E375Y/K512Y/T368F
T372Q/E375Y/K512Y/T368F
T372L/E375Y/K512Y/T368F
T372Y/K478Y/E375Y/K512Y/T368F
I370W/E375Y/K512Y/T368F
F198W/E375Y/K512Y/T368F
L381A/E375Y/K512Y/T368F
E375Y/K512Y/T368F

[0122] The polymerase reaction conditions can also be important for obtaining a two slow-step enzyme system. In particular, polymerase reaction conditions include components selected to produce two slow-step kinetics. The polymerase reaction conditions include the type and concentration of buffer, the pH of the reaction, the temperature, the type and concentration of salts, the presence of particular additives which influence the kinetics of the enzyme, and the type, concentration, and relative amounts of various cofactors, including metal cofactors. The term "polymerase reaction conditions" as used herein generally excludes the concentration of the polymerase enzyme or the concentration of the primer-template complex. Thus, two reactions are run under substantially the same polymerase reaction conditions where the first reaction has a small amount of polymerase enzyme, such as a single polymerase enzyme, and a small amount of primer template complex, such as a single primer-template complex associated with a single polymerase enzyme, and the second reaction has a higher concentration of polymerase enzyme, for example a concentration of polymerase enzyme of about 0.05 μM to 0.5 μM, and about 0.01 μM to about 0.1 μM.

[0123] It some embodiments the type and concentration of buffer are chosen in order to produce a reaction having two slow steps. Enzymatic reactions are often run in the presence of a buffer, which is used, in part, to control the pH of the reaction mixture. We have found that

in some cases the type of buffer can influence the kinetics of the polymerase reaction in a way that can lead to two slow-step kinetics. For example, in some cases, we have found that the use of TRIS as buffer is useful for obtaining a two slow-step reaction. Buffers suitable for the invention include, for example, TAPS (3-{[tris(hydroxymethyl)methyl]amino}propanesulfonic acid), Bicine (N,N-bis(2-hydroxyethyl)glycine), TRIS (tris(hydroxymethyl)methylamine), ACES (N-(2-Acetamido)-2-aminoethanesulfonic acid), Tricine (N-tris(hydroxymethyl)methylglycine), HEPES 4-2-hydroxyethyl-1-piperazineethanesulfonic acid), TES (2-{[tris(hydroxymethyl)methyl]amino}ethanesulfonic acid), MOPS (3-(N-morpholino)propanesulfonic acid), PIPES (piperazine-N,N'-bis(2-ethanesulfonic acid)), and MES (2-(N-morpholino)ethanesulfonic acid).

[0124] The pH of the reaction can influence the kinetics of the polymerase reaction, and can be used as one of the polymerase reaction conditions to obtain a reaction exhibiting two slow-step kinetics. The pH can be adjusted to a value that produces a two slow-step reaction mechanism. The pH is generally between about 6 and about 9. In some cases, the pH is between about 6.5 and about 8.0. For example, the pH can be by way of illustrations between about 6.5 and 7.5. The pH can be about 6.5, 6.6, 6.7, 6.8, 6.9, 7.0, 7.1, 7.2, 7.3, 7.4, or 7.5.

[0125] The temperature of the reaction can be adjusted in order to obtain a reaction exhibiting two slow-step kinetics. The reaction temperature may depend upon the type of polymerase which is employed. Temperatures between about 15^{O} C and 90^{O} C, between about 20^{O} C and 50^{O} C, between about 20^{O} C and 40^{O} C, or between about 20^{O} C and 30^{O} C can be used.

[0126] In one aspect, the present invention is directed to the use of a mixture of catalytic and non-catalytic metal ions in a nucleic acid synthesis reaction, to modulate the reaction kinetics of the complex. Thus, in at least one aspect, the invention is directed to nucleic acid synthesis reaction mixtures that include both catalytic and non-catalytic metals. The molar ratio of catalytic to non-catalytic metals in the reaction mixture will generally vary depending upon the type of kinetic modulation desired for a given synthesis reaction, where slower incorporation would suggest higher levels of non-catalytic metal ions. Typically, such ratios of catalytic to non-catalytic metals in the reaction mixture will vary from about 10:1 to

about 1:10, and preferably, from about 10:1 to about 1:5, depending upon the desired level of modulation, the particular enzyme system employed, the catalytic and non-catalytic metal cofactors that are used, and the reaction conditions. In particularly preferred aspects, the ratios of catalytic to non-catalytic metals will be in the range of from about 5:1 to about 1:1, with ratios of from about 2.5:1 to about 1.5:1 being particularly preferred.

[0127] In addition to the presence of such metals at the ratios described herein, the absolute concentration of such metals in the reaction mixtures will typically range from about 0.05 mM to about 50 mM, in some cases from about 0.1 mM to about 10 mM, in some cases from about 0.1 mM to about 5 mM. The composition can include, for example, from about 0.1 mM $MnCl_2$ to about 1 mM $MnCl_2$ and from about 0.1 mM $CaCl_2$ to about 2 mM $CaCl_2$; or from about 0.2 mM $MnCl_2$ to about 1 mM $MnCl_2$ and from about 0.4 mM $CaCl_2$ to about 1.5 mM $CaCl_2$.

[0128] As will be appreciated, the two slow-step enzyme systems described above can be used in combination with any other embodiments described herein, including: (1) reaction conditions that increase the frequency of branching fraction nonincorporation events; (2) modified recombinant polymerases that exhibit increased branching fractions as compared to the corresponding wild-type polymerases; (3) iterative sampling of unincorporatable nucleotides; (4) detection of noncognate branching events; (5) modified recombinant polymerases that exhibit altered translocation properties as compared to the corresponding wild-type polymerases; (6) modified recombinant polymerases that exhibit increased nucleotide residence time; (7) sequencing of nucleic acid templates using more than one mode; and/or (8) any other combination of embodiments described herein.

E. Enhanced Sequencing By Detection of Noncognate Branching Events

[0129] Nucleic acid sequencing approaches that utilize fluorescent-labeled nucleotide analogs typically require the detection of at least four colors – one color for each nucleotide representing the four different bases A, T/U, C, or G. Sequencing by incorporation methods, e.g., real time single molecule sequencing, also typically involves the detection of at least four different fluorescent labels corresponding to each of four nucleotide analogs. New approaches for reducing the number of different fluorescent labels to be detected by a sequencing system

are provided herein; these approaches can reduce the cost of nucleic acid sequencing by reducing, e.g., the number of color channels and other associated optical capabilities of the detection system.

[0130] The present invention provides new methods for nucleic acid sequencing that reduce the number of required colors to be detected by the sequencing system. Such methods utilize noncognate branching to determine which nucleotide is incorporated at a particular incorporation site during a nucleic acid polymerization reaction. For the purposes of this disclosure, noncognate branching refers to the branching (or iterative sampling) of a nucleotide, where the base of the nucleotide would incorrectly pair (not form a correct Watson-Crick base pair) with the base at the incorporation site of the template nucleic acid. For example, when the incorporation site of a template nucleic acid contains the base guanine (or G), it has been observed that branching (or sampling) of dGTP nucleotides or nucleotide analogs occurs at an appreciable frequency. The frequency of noncognate branching is sufficient to permit the identification of the nucleotide actually incorporated, regardless of whether the nucleotide actually incorporated is labeled. For instance, multiple signal pulses derived from noncognate branching (i.e., iterative sampling) of fluorescently-labeled G-containing nucleotides or nucleotide analogs can be utilized by a signal detection and sequencing system to determine that a C-containing nucleotide (which does not require a fluorescent label) was actually incorporated at that site. Thus, the temporal sequence of incorporation of four different nucleotides can be ascertained by utilizing, at most, three different fluorescent labels.

[0131] Noncognate branching of T-containing nucleotides has also been observed when the base at the incorporation site of the template is T. Accordingly, it is possible to determine the sequence of a nucleic acid template using a two-color detection system, where noncognate branching of G- and T-containing nucleotides – alone – permits the identification of incorporation of C- and A-containing nucleotides, thereby eliminating the requirement that C- and A-containing nucleotides be labeled. As will be appreciated, the present invention provides enhanced sequencing using noncognate branching, where the sampling noncognate nucleotide can be any nucleotide that would incorrectly pair with the corresponding base of the template nucleic acid.

[0132] It will be appreciated that noncognate branching events, in conjunction with a sequencing system that detects and accounts for noncognate branching events, are useful for sequencing applications other than those designed to reduce, e.g., the number of color channels required of the sequencing system. For example, even where sequencing is performed using four colors, noncognate branching can be used to validate or assist in the determination by a sequencing system of which nucleotide was incorporated at a particular site. Thus, when the base at the incorporation site of a template of unknown sequence is G, and the quantity or quality of signal pulses generated from sampling of dCTP analogs does not unambiguously permit assignment of dCTP as the incorporating nucleotide at that site, noncognate branching of labeled dGTP nucleotides facilitates the determination that dCTP was indeed incorporated at the site.

[0133] As will be appreciated, enhanced sequencing by detection and analysis of noncognate branching events can be used in combination with any other embodiments described herein, including: (1) reaction conditions that increase the frequency of branching fraction nonincorporation events; (2) modified recombinant polymerases that exhibit increased branching fractions as compared to the corresponding wild-type polymerases; (3) iterative sampling of unincorporatable nucleotides; (4) two slow-step enzyme systems; (5) modified recombinant polymerases that exhibit altered translocation properties as compared to the corresponding wild-type polymerases; (6) modified recombinant polymerases that exhibit increased nucleotide residence time; (7) sequencing of nucleic acid templates using more than one mode; and/or (8) any other combination of embodiments described herein.

II. SEQUENCING BY INCORPORATION USING MORE THAN ONE MODE OF SEQUENCING

[0134] A sequencing composition may be especially advantageous for one particular purpose or type of template nucleic acid, but may exhibit limitations for a second purpose or type of template nucleic acid. As described in detail herein, sequencing under conditions that promote a relatively high branching fraction, in combination with appropriate signal analysis, reduces the likelihood of base calling errors that result from undetected signal pulses generated from genuine nucleotide incorporation events. Accordingly, the base calling accuracy of

sequencing by incorporation reactions that exhibit a relatively high level of branch fraction nonincorporation events – when combined with a sequencing system that accounts for such redundant signal pulses per incorporation event – can be higher than the level of accuracy achieved under low branch rate conditions. On the other hand, sequencing under conditions in which the branching fraction is low typically results in improved sequence read length, as the number of incorporation events will be greater before the polymerase has an opportunity to dissociate from the template nucleic acid. Longer read lengths simplify assembly of contig information, e.g., to facilitate genomic sequencing.

[0135] De novo sequencing (sequencing a template nucleic acid of unknown sequence, e.g., genomic DNA of unknown sequence) is, in some cases, optimally performed under a combination of both high accuracy and high read length conditions. Identifying one particular sequencing composition that adequately fulfills both of these competing requirements is not necessary in the present invention. Instead, the present invention provides new methods and systems for sequencing a template nucleic acid, in which the template nucleic acid is sequenced using more than one sequencing mode ("multi-modal" or "variable mode" sequencing). As used herein, a sequencing mode refers to a sequencing composition (e.g., a mixture of a particular nucleic acid polymerase, nucleotides or nucleotide analogs, metal cofactors, and other components of a sequencing reaction) and other conditions that affect nucleic acid polymerization, e.g., reaction temperature. The methods and systems of the invention utilize a first sequencing mode that confers a particular benefit for purposes of generating a sequence readout (e.g., high accuracy), and then switch to a second (or more) sequencing mode that confers a benefit not realized during the first sequencing mode, e.g., a high sequence read length. As will be appreciated, any number of different modes can be employed by the methods and systems of the inventions, e.g., 2, 3, 4, 5, 10, or 20 different modes may be used to sequence a particular nucleic acid template until the desired results are achieved. For example, a template nucleic acid can be sequenced multiple times using a "high accuracy" mode and/or multiple times using a "high read length" mode, in order to obtain the accuracy and fold coverage useful for, e.g., de novo sequencing. Alternatively, a single template nucleic acid can be partially sequenced using, e.g., a high accuracy mode, and the mode can be switched prior to complete sequencing of the template to, e.g., a high read length mode.

[0136] A sequencing mode in which, e.g., high accuracy is desirable can employ, e.g., reaction conditions that induce a relatively high average branching fraction during the nucleic acid polymerization reaction. Such reaction conditions are described in detail herein. In one embodiment of the invention, one or more metal cofactors may be included in the sequencing composition at concentrations and/or ratios that induce a relatively high average branching fraction. Metal cofactors of the invention include, e.g., Mg^{++}, Mn^{++}, Zn^{++}, Co^{++}, Ca^{++}, Fe^{++}, Cr^{++}, and/or Sr^{++}. Reaction conditions in which metal cofactor identities, concentrations and/or ratios induce high levels of branching are described in detail herein. As described herein, a relatively high level of branching fraction nonincorporation events can be induced in a sequencing composition that includes, e.g., a relatively low concentration of Mn^{++}, a sequencing composition that includes Mg^{++}, a sequencing composition that includes both Mg^{++} and Mn^{++} with Mg^{++} being included at a higher concentration than Mn^{++}, and a sequencing composition that includes Mn^{++} and a metal cofactor other than Mg^{++}, e.g., calcium, barium, strontium, iron, cobalt, nickel, tin, zinc, and europium.

[0137] A sequencing mode in which, e.g., high read length is desirable can employ, e.g., reaction conditions that discourage branch fraction nonincorporation events during the nucleic acid polymerization reaction. Such reaction conditions are described herein. As described herein, reaction conditions that do not promote branching events include, e.g., Mn^{++} alone at a concentration at which branching is not induced (e.g., greater than 250 μM), or Mn^{++} in combination with a second metal cofactor, e.g., Mg^{++}, with the concentration of Mn^{++} being greater than the concentration of Mg^{++}. Desirably, sequencing modes of the present invention that produce high read lengths (e.g., a high number of successive incorporation events that are detected by sequencing systems of the invention) can produce read lengths of preferably more than 200 base pairs (bp) or more, more preferably 500 bp or more, more preferably 1000 bp or more, more preferably 10,000 bp or more, or more preferably 50,000 bp or more.

[0138] In order to permit the switching between a first sequencing mode and a second (or subsequent or more) sequencing mode, the sequencing composition – including the template nucleic acid and polymerase – is desirably confined within a structure to which additional components can be added to the sequencing composition to achieve the desired subsequent mode. In one aspect, for example, the sequencing of a template nucleic acid under

a first sequencing mode occurs within a structure to which a channel, e.g., a microfluidic channel, delivers the reagents necessary to achieve the subsequent mode. For example, switching from a high branch fraction sequencing mode, in which Mg^{++} is the sole catalytic metal ion, to a high read length sequencing mode can be achieved by delivery of Mn^{++} to the sequencing reaction, such that the final concentration of Mn^{++} is about 300 μM or more, about 400 μM or more, about 500 μM or more, about 700 μM or more, or about 1 mM or more.

[0139] In a preferred embodiment, the sequencing reactions take place in a structure that provides optical confinement, e.g., a nanohole or zero-mode waveguide. Further details regarding confinement strategies, substrates and systems for monitoring sequencing reactions can be found in co-pending published U.S. Patent Application No. 2007-0188750, and published International Patent Application No. WO 2007/095119, the full disclosures of which are incorporated herein by reference in their entirety for all purposes.

[0140] The sequencing methods and systems of the present invention can employ a wide variety of template nucleic acids. In some cases, the template nucleic acid is a linear template. For a linear template, switching between a high accuracy and a high read length mode can result in a read wherein a relatively long stretch of template nucleic acid is sequenced in a given time, wherein generally, multiple regions along the length will be sequenced with high accuracy. The knowledge that the high accuracy regions are arranged in a particular sequential manner can be useful in putting together the sequence of the nucleic acid. In some cases, the template nucleic acid is a circular template. For a circular template, switching between high accuracy and long read length modes can result in the same region of the template nucleic acid sequenced by each of the modes. In a preferred embodiment, the template to be sequenced in more than one sequencing mode is a single-stranded nucleic acid loop. Double-stranded templates can reanneal, reducing primer annealing efficiency and impeding the polymerase-catalyzed extension of a sequencing reaction. In fact, loops can also be preferable to linear templates because a DNA polymerase can only copy a linear template, e.g., to which a primer has been annealed, once before it falls off the distal end of the template. In contrast, a strand-displacing polymerase can replicate a contiguous nucleic acid loop several times. The primer that is annealed to the loop is eventually displaced at its 5'-end upon completion of one revolution of the polymerase around the nucleic acid loop, and as

polymerization and displacement continue, a linear, single-stranded product comprising several copies of the nucleic acid sequence of the loop is generated. Accordingly, using nucleic acid loops in sequencing can provide an internal sequencing control.

[0141] The methods for preparing closed, single-stranded nucleic acid loops include providing a genomic DNA, a cDNA, or a DNA concatamer and generating double-stranded fragments that each comprise a first strand (e.g., an exonuclease sensitive strand) and a second strand (e.g., an exonuclease resistant strand). In a following step, the two strands in each fragment are separated, and the resulting single-stranded fragments are circularized to produce closed single-stranded nucleic acid loops, which can then be used as templates in a high-throughput sequencing system. Further details regarding the preparation of single stranded nucleic acid loops, and their use as templates in high-throughput sequencing systems, can be found in Patel et al. PCT Application Serial No. PCT/US2009/001930 METHODS AND COMPOSITIONS FOR NUCLEIC ACID SAMPLE PREPARATION, incorporated herein by reference in its entirety for all purposes.

[0142] Systems for analyzing the data generated during sequencing of a template nucleic acid using more than one sequencing mode are also a feature of the invention. Such systems will include a signal detector, e.g., in the case of a plurality of arrayed sequencing reactions, an array detector, e.g., an EMCCD. The detector, is then operatively coupled to a data storage and processing system. In a first sequencing mode, e.g., a mode where a high branching fraction is induced, the processing system is capable of interpreting multiple, e.g.,. redundant, or iterative signal pulses for each nucleotide incorporation event during a sequencing reaction to call bases with increased accuracy. In a subsequent sequencing mode, e.g., a high read length mode in which little or no branching occurs, the processing system is capable of interpreting single signal pulses as nucleotide incorporation events. Further details regarding base calling during sequencing by incorporation methods are found in Tomaney et al. PCT Application Serial No. PCT/US2008/065996 METHODS AND PROCESSES FOR CALLING BASES IN SEQUENCING BY INCORPORATION METHODS, incorporated herein by reference in its entirety for all purposes.

[0143] The multi-modal sequencing method of the present invention optionally includes recording signals or signatures, e.g., in a computer readable medium. The signals or signatures can be stored, e.g., as graphic or digital information. Any typical recording device such as a hard drive, memory card, memory stick, optical storage device or floppy drive can be used to record detected signals. Signals or signal signatures can also be deconvoluted or translated to provide, e.g., sequence information, e.g., in sequencing systems of the invention. Signal processing equipment can include, e.g., a computer having appropriate software for converting signals into sequence or assay parameter information.

[0144] Signal detection optics can be coupled to cameras, digital processing apparatus, or the like, to record and analyze signals detected in the various systems herein, . Systems can include a microscope, a CCD, a phototube, a photodiode, an LCD, a scintillation counter, film for recording signals, and the like. A variety of commercially available peripheral equipment and software is available for digitizing, storing and analyzing a digitized video or digitized optical image, e.g., using PC (Intel x86 or pentium chip-compatible DOS™, OS2™ WINDOWS™, WINDOWS NT™ or WINDOWS95™ based machines), MACINTOSH™, LINUX, or UNIX based (e.g., SUN™ work station) computers or digital appliances. Computers and digital appliances can include software for analyzing and perfecting signal interpretation. This can include standard application software such as spreadsheet or database software for storing signal information. However, systems of the invention can also include statistical analysis software to interpret signal or signature data, e.g., to translate the data into nucleic acid sequence information. For example, many vendors, such as Partek Incorporated (St. Peters, Mo.; www.partek.com) provide software for pattern recognition which can be applied to signal interpretation and analysis. Algorithms for sequencing systems that can be adapted to the invention are also described in Tomaney et al. PCT Application Serial No. PCT/US2008/065996 METHODS AND PROCESSES FOR CALLING BASES IN SEQUENCING BY INCORPORATION METHODS, incorporated herein by reference in its entirety for all purposes. Once signal information has been converted into sequence information, standard sequence aalysis software can be used to assemble overlapping sequence information. For example, sequence contigs can be assembled using available software such as DNA Baser (Heracle Software, Germany), or Artemis 11 (Sanger Institute) "Artemis and ACT:

Viewing, annotating and comparing sequences stored in a relational database" Carver et al. Bioinformatics 2008 PMID: 18845581 DOI: 10.1093/bioinformatics/btn529).

[0145] Relationships between datasets (e.g., high accuracy versus high readlength data) can similarly be analyzed, e.g., by pattern recognition software, Bayes classifiers, neural networks, Monte Carlo analysis, Principal Component Analysis (PCA), etc. Further information regarding genetic algorithms and neural networks that can be used to analyze signal or signature information can be found in David E. Goldberg (1989) Genetic Algorithms in Search, Optimization and Machine Learning Addison-Wesley Pub Co; ISBN: 0201157675; Timothy Masters (1993) Practical Neural Network Recipes in C++(Book & Disk edition) Academic Pr; ISBN: 0124790402; Kevin Gurney (1999) An Introduction to Neural Networks, UCL Press, 1 Gunpowder Square, London EC4A 3DE, UK; Christopher M. Bishop (1995) Neural Networks for Pattern Recognition Oxford Univ Press; ISBN: 0198538642; Brian D. Ripley, N. L. Hjort (Contributor) (1995) Pattern Recognition and Neural Networks Cambridge Univ Pr (Short); ISBN: 0521460867;Rubinstein, R. Y.; Kroese, D. P. (2007) Simulation and the Monte Carlo Method (2nd ed.). New York: John Wiley & Sons. ISBN 9780470177938; Tarantola, Albert (2005) Inverse Problem Theory Philadelphia: Society for Industrial and Applied Mathematics ISBN 0898715725; Steeb (2008) The Nonlinear Workbook: Chaos, Fractals, Neural Networks, Genetic Algorithms, Gene Expression Programming, Support Vector Machine, Wavelets, Hidden Markov Models, Fuzzy Logic with C++, Java and SymbolicC++ Programs: 4th edition. World Scientific Publishing. ISBN 981-281-852-9; Sergios Theodoridis, Konstantinos Koutroumbas, (2009) Pattern Recognition (4th edition), Elsevier, ISBN 978-1-59749-272-0, and in a variety of other currently available references. Computers/digital appliances also optionally include or are operably coupled to user viewable display systems (monitors, CRTs, printouts, etc.), printers to print data relating to signal information, peripherals such as magnetic or optical storage drives, user input devices (keyboards, microphones, pointing devices) and the like.

[0146] One example of a system that utilizes more than one mode of sequencing is illustrated in **Figure 7**. As shown, system 700 includes substrate 702 of the invention, e.g., upon which sequencing reactions are performed. Optical energy source 704 and additional optical energy source 706 deliver excitation light to the substrate 702, via an optical train. As

shown, the optical train includes a dichroic mirror 708 that is transmissive to excitation wavelength of the light from laser 704, while being reflective of light of the wavelength produced by laser 706, allowing both illumination beams to be in the same path. The excitation light is then directed at a second dichroic mirror 710, that reflects the excitation light through objective lens 712, and onto the substrate 702. Optical signals from the substrate 702 are then collected by objective lens 712, and are passed through dichroic 710. The fluorescent signals are then subjected to spatial separation using, e.g., a dispersive optical element, such as an optical grating or prism 714. The separated signals are then focused upon array detector 716, e.g., an EMCCD. The detector, is then operatively coupled to a data storage and processing system, such as computer 718 for processing and storage of the signal data and presentation of the data in a user desired format, e.g., on printer 720.

[0147] As shown in Figure 7, in one example of multi-modal sequencing, a first sequencing mode (Panel I) is produced in which a sequencing by incorporation reaction occurs under higher branch fraction conditions than a second sequencing mode. During the first sequencing mode, the processing system of the computer is capable of interpreting multiple, e.g.,. redundant, or iterative signal pulses or signatures for each nucleotide incorporation event during a sequencing reaction to call bases. Subsequently, a second sequencing mode (Panel II) is produced in which a sequencing by incorporation reaction occurs with no or very few branch fraction nonincorporation events, and accordingly, longer readlengths. During the second sequencing mode, the processing system of the computer interprets one or very few signal pulses or signatures as a nucleotide incorporation event during the sequencing reaction to call bases. The system can utilize signal pulse or signature data from each sequencing mode in order to determine the sequence of the template nucleic acid. As will be appreciated, a portion of the nucleic acid template can be sequenced once (in one of the two or more sequencing modes) or two or more times (two or more times in the same or multiple sequencing modes).

[0148] Although illustrated as an optical train that is transmissive of fluorescent signals, e.g., as provided by dichroic **710**, it will be appreciated that fluorescence reflective systems may also be employed. Further details regarding base calling during sequencing by incorporation methods are found in Tomaney et al. PCT Application Serial No. PCT/US2008/065996 METHODS AND PROCESSES FOR CALLING BASES IN

SEQUENCING BY INCORPORATION METHODS, incorporated herein by reference in its entirety for all purposes.

[0149] As will be appreciated, sequencing of a nucleic acid template in more than one sequencing mode can be combined with (1) reaction conditions that increase the frequency of branching fraction nonincorporation events; (2) modified recombinant polymerases that exhibit increased branching fractions as compared to the corresponding wild-type polymerases; (3) iterative sampling of unincorporatable nucleotides; (4) two slow-step enzyme systems; (5) modified recombinant polymerases that exhibit altered translocation properties as compared to the corresponding wild-type polymerases; (6) modified recombinant polymerases that exhibit increased nucleotide residence time as compared to the corresponding wild-type polymerases; and/or (7) any other combination of embodiments described herein.

III. DELAYED TRANSLOCATION

[0150] The present invention also provides polymerases that exhibit delayed translocation as compared to parental/wild type enzymes. Time necessarily lapses between the incorporation of one nucleotide and the incorporation of the next nucleotide due to the sequential (rather than simultaneous) nature of nucleotide incorporation. The duration of this time lapse is determined primarily by the rate at which a polymerase translocates along a template polynucleotide between incorporation events ("translocation" refers to the movement of a DNA polymerase along a template polynucleotide from an initial enzyme binding site to a subsequent enzyme binding site, where the enzyme binding sites correlate to nucleotide incorporation sites). Upon incorporation of a nucleotide, a polymerase is unable to accept another nucleotide until it has gone through the translocation process and moved into the next incorporation site.

[0151] When a polymerase translocates at a typical wild-type rate from one incorporation site to the next, a single molecule sequencing (SMS) system has less time to identify and distinguish when the polymerase has moved into the next incorporation site because the rapid translocation affords little separation between the signal pulses arising from incorporation events at one incorporation site and the pulses arising from those events at the next incorporation site. Accordingly, SMS under conditions of typical polymerase

translocation rates can have an increased potential for deletion errors relative to the correct template sequence (i.e., signals from incorporation events may not be detected or may be detected but not interpreted as incorporation events). These deletions would be artifacts that constitute errors in sequencing reads.

[0152] In accordance with aspects of the present invention, however, the modified recombinant polymerases provided by the present invention that exhibit a translocation delay of longer duration as compared to parental (or wild-type) polymerases are advantageous in the context of SMS because they provide greater temporal separation (i.e., resolution) between the signal pulse emitted while the polymerase resides at one incorporation site from the pulse emitted once the polymerase has entered the next incorporation site. This increased translocation delay permits the sequencing system to interpret signal pulses as correlating to separate and distinct nucleotide incorporation events, and hence greatly diminishes the potential for missed pulses or inaccurately characterized pulses relative to the correct template sequence. Further, a delay in translocation enhances the ability of a sequencing system to distinguish branching events that exhibit short inter-pulse widths from incorporation events that, under the conditions of delayed translocation, exhibit markedly broader inter-pulse widths.

[0153] In one aspect, the polymerases of the present invention are engineered to exhibit two sequential translocation kinetic steps that occur at a slower rate as compared to a parental polymerase, e.g., a wild-type parental polymerase. The two translocation kinetic steps may occur at the same rate, or the ratio of their rates may vary up to about 1:5 or 5:1 or more. The occurrence of exceedingly short inter-pulse widths, e.g., the time delay between signal pulses derived from sequential nucleotide incorporation events, is markedly reduced when two slow translocation kinetic steps are engineered into the polymerase. However, the present invention also provides polymerases that exhibit a translocation delay arising from a slow translocation kinetic step in conjunction with a slow kinetic step not related to translocation, e.g., a slow nucleotide and/or nucleotide analog binding step. Further details regarding polymerase systems with two slow kinetic steps can be found in Bjornson et al. PCT Application Serial Number PCT/US2009/002003 TWO SLOW-STEP POLYMERASE ENZYME SYSTEMS AND METHODS, incorporated herein by reference in its entirety for all purposes.

[0154] The delay in translocation exhibited by the polymerases of the invention can be more than 2.5x greater, e.g., more than 5x greater, more than 10x greater, more than 15x greater, more than 50x greater, more than 100x greater, more than 1000x greater, or more than 10,000x greater as compared to a parental polymerase (e.g., a wild type Φ29 polymerase) or more. Modified recombinant polymerases of the invention can allow the translocation step of a template-dependent polymerization reaction to be observable, e.g., wherein an otherwise unobservable translocation step (e.g., due to a high rate of translocation of an unmodified polymerase and limitations of the detection system) becomes observable as a result of modification of the polymerase.

[0155] The present invention also provides modified recombinant polymerases that include a heterologous polypeptide sequence fused at or near the c-terminus of the polymerase and/or amino acid substitutions or deletions, in order to delay translocation of the polymerase. As noted above, when a polymerase translocates at a typical wild-type rate from one incorporation site to the next, a single molecule sequencing (SMS) system can fail to distinguish when the polymerase has moved into the next incorporation site because the rapid translocation affords little separation between the signal pulses arising from branching and incorporation events at one incorporation site and the pulses arising from those events at the next incorporation site. Accordingly, SMS under conditions of typical polymerase translocation rates is susceptible to incorrect insertions relative to the correct template sequence (i.e., multiple signal pulses arising from branching and incorporation events at a single incorporation site may be interpreted as pulses arising from incorporation events at more than one incorporation site). These insertions are artifacts that constitute errors in sequencing reads.

[0156] The present invention provides modified or recombinant DNA polymerases that exhibit decreased translocation rates in order to provide greater temporal separation (i.e., resolution) between the signal pulses emitted while the polymerase resides at one incorporation site from the pulses emitted once the polymerase has entered the next incorporation site. This decreased translocation rate permits the sequencing system to interpret clusters of signal pulses as correlating to nucleotide incorporation sites, and hence greatly diminishes the occurrence of incorrect insertions relative to the correct template sequence.

[0157] In one embodiment, delayed translocation is accomplished by encoding a Φ29 DNA polymerase with a heterologous polypeptide sequence fused to the c-terminus of the DNA polymerase. For the purposes of this disclosure, "heterologous" refers to a polypeptide sequence that is not present in the wild-type parental polymerase. That these polymerases retain their functionality is a surprising aspect of the invention. The active site of the polymerase is located in the c-terminal portion of the protein, and previous attempts to modify the c-terminal portion have rendered the polymerase inactive. In one aspect, the heterologous polypeptide sequence can include between 6 and 10 positively charged amino acids, e.g., histidine. This stretch of positively charged amino acids can be encoded immediately downstream of the c-terminus of the polymerase. An example polymerase of this embodiment, comprising a Φ29 polymerase fused to 10 histidine residues at its c-terminus (SEQ ID No. 1), exhibits a translocation rate of 84 sec^{-1}, which constitutes a 2.3x delay in translocation as compared to an unmodified parental Φ29 polymerase under identical conditions. In another aspect, an amino acid linker sequence, e.g., a Ser3Gly linker (e.g., Gly-Gly-Gly-Ser-Gly-Gly-Gly-Ser-Gly-Gly-Gly-Ser-Gly) is encoded between the stretch of positively charged amino acids and the C-terminus of the polymerase. An example polymerase of this embodiment, comprising a Φ29 polymerase fused to 10 histidine residues at its c-terminus, with a Ser3Gly linker between the polymerase and the histidine residues (SEQ ID No. 2), exhibits a translocation rate of 220 sec^{-1}, which constitutes a 2.4x delay in translocation as compared to an unmodified parental Φ29 polymerase under identical conditions. The affinity of positively charged amino acid residues at the c-terminus of the polymerase for the negatively charged phosphate backbone of DNA decreases the efficiency of polymerase translocation to the next nucleotide incorporation site, thereby causing a delay in translocation. Polymerases of the invention can also exhibit improved polymerase processivity, as the interaction between the polymerase and the phosphate backbone of DNA is enhanced by, e.g., the addition of positively-charged amino acid residues at the c-terminus.

[0158] In another embodiment, the polymerases that exhibit delayed translocation comprise amino acid substitutions and/or deletions that modulate interaction of the polymerases with the negatively-charged phosphate backbone of DNA. A number of specific examples of a modified polymerase, e.g. modified to delay polymerase translocation, are described herein. A

region of the polymerase responsible for interaction with the DNA template and primer is referred to herein as the binding cleft. Within the binding cleft are particular amino acid residues that interact with the phosphate backbone of DNA. The phosphate backbone constitutes a uniform interaction platform, e.g., nucleobase independent, and interactions with the phosphate backbone are altered in polymerases of the present invention, e.g., utilizing the negative charge of the backbone. For example, the interaction between a polymerase of the invention and the phosphate backbone of DNA can be enhanced by substituting neutral or negatively charged amino acids in the binding cleft with positively charged residues. However, mutations that delay polymerase translocation are not limited to this region of the polymerase. Relative to a wild-type Φ29 DNA polymerase, polymerase modifications of the present invention can include, e.g., any of the following mutations or combination of the following mutations: Asp570Lys; Asp570Ala; Asn313Lys; Asn313Ala; Gln303Lys; Gln303Ala; Gly532Ser; Met533delet; Cys530delet; Met533delet and Cys530delet; Gly532delet; Ala531Gly; Gly532Ser; Thr573Lys; Thr573Ala; Asn396Lys; Thr571Lys; Thr571Ala; Thr534Lys; Thr534Ala; Asp535Lys; Asp534Ala; Lys529Ala; and Lys529Asn. For the purposes of the present application, a mutation that includes a deletion at a particular residue position is presented by the amino acid abbreviation, followed by the residue position, followed by "del". Thus, for example, the Met533delet mutation presented above will be understood to mean a mutant polymerase in which the methionine at position 533 has been deleted.

[0159] The translocation delay exhibited by modified polymerases of the present invention, e.g., a modified Φ29 polymerase or a modified Φ 29-type polymerase, can be greater than the translocation delay exhibited by the parental polymerase or, e.g., about 2.5x greater, about 5x greater, about 10x greater, about 15x greater, about 50x greater or about 100x or more greater, as compared to a parental polymerase (e.g., a wild type Φ29 polymerase) under the standard conditions described above.

[0160] As will be appreciated, the above-identified modified or recombinant polymerases that display delayed translocation may optionally include additional modifications that confer other useful properties described herein, e.g., increased branching fraction, increased nucleotide analog residence time and/or increased processivity, etc. As will also be

appreciated, the above-identified modified or recombinant polymerases that display delayed translocation can be used in combination with any other embodiments described herein, including: (1) reaction conditions that increase the frequency of branching fraction nonincorporation events; (2) modified recombinant polymerases that exhibit increased branching fractions as compared to the corresponding wild-type polymerases; (3) iterative sampling of unincorporatable nucleotides; (4) two slow-step enzyme systems; (5) detection of noncognate branching events; (6) modified recombinant polymerases that exhibit increased nucleotide residence time; (7) sequencing of nucleic acid templates using more than one mode; and/or (8) any other combination of embodiments described herein.

IV. INCREASED RESIDENCE TIME

[0161] The present invention also provides modified recombinant polymerases that exhibit increased nucleotide or nucleotide analog residence time at an active site of the polymerase. During SMS, a number of signal parameters may be and generally are used for pulse identification. Two primary parameters are pulse intensity and pulse width, where pulse width relates to the duration of a signal pulse as detected by a signal detector of the sequencing system. Signal pulses that exhibit a short pulse width can arise from undesirable sources during SMS, e.g., transient pulses from labeled analogs in the reaction region that are not involved in an incorporation event. Incorporation events, on the other hand, are generally characterized by longer pulse durations stemming from increased residence time of the labeled nucleotide in the observation region by virtue of it being complexed with the polymerase.

[0162] Notwithstanding the differences in residence time duration upon incorporation, in some cases it would be desirable to further increase residence time for incorporated nucleotides, in order to further enhance the distinction between incorporation and transient signal events.

[0163] The modified recombinant polymerases of the present invention improve the accuracy of template-dependent polymerization reactions by increasing the residence times of nucleotides or nucleotide analogs at an active site of the polymerase, thereby producing signal pulses of increased width to further facilitate identification as signal pulses resulting from incorporation events.

[0164] As noted above, signal pulse intensity and width are primary parameters for signal pulse identification during template-dependent polymerization reactions, e.g., SMS utilizing dye-labeled nucleotide analogs. The signal pulse width is largely determined by the residence time of the nucleotide analog at the active site of the polymerase during the nucleotide incorporation step of the polymerase kinetic cycle. Insufficient residence times produce signal pulses of short duration and, accordingly, narrow pulse width that can prevent detection of genuine nucleotide incorporation events by a signal detector of the sequencing system. Undetected incorporation events arising from insufficient residence time and, accordingly, signal pulse width, constitute sequencing errors that diminish the reliability of SMS results.

[0165] The present invention provides modified or recombinant polymerases that improve sequence read accuracy by increasing the residence time of a nucleotide analog at the polymerase active site. During the polymerase kinetic cycle, interaction between a DNA polymerase and a nucleotide or nucleotide analog induces a conformational change in the polymerase, referred to herein as polymerase isomerization. During polymerase isomerization, the nucleotide or nucleotide analog is effectively immobilized at the active site of the polymerase. The duration of the isomerization step, therefore, affects the residence time of the nucleotide analog at the polymerase active site, and, accordingly, affects the signal pulse width as detected by a signal detector of a sequencing system. A slow isomerization step, therefore, can increase signal pulse width and enhance the accuracy of applications such as SMS. Further details regarding DNA polymerase enzymes and/or nucleotides or nucleotide analogs that provide altered residence times for enhanced nucleic acid sequence analysis and determination can be found in Rank et al. U.S. Application Serial No. 11/977,160 POLYMERASE ENZYMES AND REAGENTS FOR ENHANCED NUCELIC ACID SEQUENCING, incorporated herein by reference in its entirety for all purposes.

[0166] Completion of the nucleotide or nucleotide analog incorporation step of the DNA polymerase kinetic cycle requires dissociation of the polymerase from the incorporating nucleotide. Until dissociation occurs, the nucleotide, e.g., a dye-labeled nucleotide analog, is effectively immobilized at the active site of the polymerase. Slowing the dissociation step, therefore, results in increased nucleotide residence time at the active site of the polymerase and

increases the duration of a signal pulse emitted by a dye-labeled nucleotide analog as it is incorporated into the copy nucleic acid.

[0167] Polymerases of the present invention increase the residence time of nucleotides or nucleotide analogs by slowing the isomerization and/or dissociation steps of the polymerase kinetic cycle. Accordingly, these polymerases facilitate signal pulse detection by increasing signal pulse width. For example, a polymerase provided by the present invention comprises the amino acid substitution T368P and exhibits an average nucleotide residence time that is about 2x greater as compared to the residence time of a wild-type parental polymerase under identical conditions.

[0168] The nucleotide or nucleotide analog residence time exhibited by modified polymerases of the present invention, e.g., a modified Φ29 polymerase or a modified Φ 29-type polymerase, can be greater than the residence time exhibited by the parental polymerase or, e.g., about 1.5x greater, about 2.0x greater or about 3.0x or more greater, under the standard conditions described above.

[0169] As will be appreciated, the above-identified modified or recombinant polymerases that display increased nucleotide or nucleotide analog residence time may optionally include additional modifications that confer other useful properties described herein, e.g., increased branching fraction, delayed translocation and/or increased processivity. As will also be appreciated, the above-identified modified or recombinant polymerases that display increased nucleotide or nucleotide analog residence time can be used in combination with any other embodiments described herein, including: (1) reaction conditions that increase the frequency of branching fraction nonincorporation events; (2) modified recombinant polymerases that exhibit increased branching fractions as compared to the corresponding wild-type polymerases; (3) iterative sampling of unincorporatable nucleotides; (4) two slow-step enzyme systems; (5) detection of noncognate branching events; (6) modified recombinant polymerases that exhibit altered translocation properties as compared to the corresponding wild-type polymerases; (7) sequencing of nucleic acid templates using more than one mode; and/or (8) any other combination of embodiments described herein.

[0170] The properties of increased branching rates, delayed translocation and increased nucleotide or nucleotide analog residence time are particularly useful in the context of an incorporation of labeled nucleotides by the polymerase, e.g., as detected during sequencing by incorporation methods (including, e.g., SMS methods). For example, the invention provides, e.g., compositions that include one or more engineered or modified polymerase enzymes optionally with one or more template DNAs, and/or labeled or otherwise modified nucleotides or nucleotide analogs, where the composition exhibits increased branching rates and/or delayed polymerase translocation during template dependent polymerase-mediated nucleic acid synthesis. Methods, including SMS using these compositions, are also provided, as are general methods of making polymerases having the properties noted herein.

[0171] Accordingly, among other aspects, the present invention provides new polymerases that incorporate nucleotide analogs, such as phosphate labeled analogs, into a growing template copy during DNA amplification. These polymerases are modified such that they have increased branching rates and/or delayed translocation and/or increased residence time when incorporating the relevant analogs, and optionally have improved DNA-polymerase processivity as compared to corresponding wild-type parental polymerases (e.g., polymerases from which modified recombinant polymerases of the invention were derived, e.g., by mutation).

[0172] These new polymerases and reaction conditions are particularly well suited to DNA amplification and/or sequencing applications, particularly sequencing protocols that include detection in real time of the incorporation of labeled analogs into DNA amplicons, because the increased branching rate, delayed translocation and/or nucleotide residence time facilitates the correct determination of which labeled nucleotide is incorporated at a site during a template dependent polymerization reaction.

FURTHER DETAILS REGARDING SYSTEMS FOR SEQUENCING BY INCORPORATION

[0173] One example of a system for use in the present invention is illustrated in **Figure 1**. As shown, system 100 includes substrate 102 of the invention, e.g., upon which sequencing reactions are performed. Optical energy source 104 and additional optical energy source 106

deliver excitation light to the substrate 102, via an optical train. As shown, the optical train includes a dichroic mirror 108 that is transmissive to excitation wavelength of the light from laser 104, while being reflective of light of the wavelength produced by laser 106, allowing both illumination beams to be in the same path. The excitation light is then directed at a second dichroic mirror 110, that reflects the excitation light through objective lens 112, and onto the substrate 102. Optical signals from the substrate 102 are then collected by objective lens 112, and are passed through dichroic 110. The fluorescent signals are then subjected to spatial separation using, e.g., a dispersive optical element, such as an optical grating or prism 114. The separated signals are then focused upon array detector 116, e.g., an EMCCD. The detector, is then operatively coupled to a data storage and processing system, such as computer 118 for processing and storage of the signal data and presentation of the data in a user desired format, e.g., on printer 120. The processing system of the computer is capable of interpreting multiple, e.g.,. redundant, or iterative signal pulses for each nucleotide incorporation event during a sequencing reaction to call bases with increased accuracy. Although illustrated as an optical train that is transmissive of fluorescent signals, e.g., as provided by dichroic 110, it will be appreciated that fluorescence reflective systems may also be employed. Further details regarding base calling during sequencing by incorporation methods are found in Tomaney et al. PCT Application Serial No. PCT/US2008/065996 METHODS AND PROCESSES FOR CALLING BASES IN SEQUENCING BY INCORPORATION METHODS, incorporated herein by reference in its entirety for all purposes.

[0174] As will be appreciated, a number of other components may be included in the systems described herein, including optical filters for filtering background illumination or bleed-through illumination from the light sources, from the actual optical signals. Additionally, alternate optical trains may employ cascaded spectral filters in separating different spectral signal components. A monitor of the computer can display optical signal pulse patterns 122 generated under the conditions provided by the invention, e.g., increased branching, delayed polymerase translocation or increased nucleotide analog residence time. A variety of other optical configurations may additionally be employed in conjunction with the compositions of the invention.

[0175] In the context of nucleic acid sequencing methods, it will be appreciated that the signal sources each represent sequencing reactions, and particularly, polymerase-mediated, template-dependent primer extension reactions, where in preferred aspects, each base incorporation event results in a prolonged illumination (or localization) of one of four differentially labeled nucleotides being incorporated, so as to yield a recognizable pulse or signature that carries a distinguishable spectral profile or color.

[0176] In the context of the present invention, a signal event is first classified as to whether it constitutes a significant signal pulse or signature based upon whether such signal event meets any of a number of different criteria. Once identified or classified as a significant pulse or signature, the signal pulse or signature may be further assessed to determine whether the signal pulse or signature constitutes an incorporation event and may be called as a particular incorporated base. As will be appreciated, the basis for calling a particular signal event as a significant pulse or signature, and ultimately as an incorporation event, will be subject to a certain amount of error, based upon a variety of parameters as generally set forth herein. The reaction conditions and modified recombinant polymerases of the present invention diminish the error associated with base calling.

[0177] Once a particular signal is identified as a significant pulse or signature and is assigned a particular spectrum, e.g. color, the spectrally assigned pulse may be further assessed to determine whether the pulse or signature can be called an incorporation event and, as a result, call the base incorporated in the nascent strand, or its complement in the template sequence. Calling of bases from color assigned pulse or signature data will typically employ tests that identify the confidence level with which a base is called. Typically, such tests will take into account the data environment in which a signal was received, including a number of the same data parameters used in identifying significant pulses, etc. For example, such tests may include considerations of background signal levels, adjacent pulse signal parameters (spacing, intensity, duration, etc.), spectral image resolution, and a variety of other parameters. Such data may be used to assign a score to a given base call for a color assigned signal pulse or signature, where such scores are correlative of a probability that the base called is incorrect, e.g., 1 in 100 (99% accurate), 1 in 1000 (99.9% accurate), 1 in 10,000 (99.99% accurate), 1 in 100,000 (99.999% accurate), or even greater.

[0178] Once a base is called with sufficient accuracy, subsequent bases called in the same sequencing run, and in the same primer extension reaction, may then be appended to each previously called base to provide a sequence of bases in the overall sequence of the template or nascent strand. Iterative processing and further data processing, as described in greater detail below, can be used to fill in any blanks, correct any erroneously called bases, or the like for a given sequence.

POLYMERASES AND NUCLEOTIDE ANALOGS

[0179] Various polymerases may be used in the methods, compositions and systems described herein, including DNA polymerases, RNA polymerases, reverse transcriptases, and mutant or altered forms of any of the foregoing.

[0180] DNA polymerases that can be modified to increase the average branching fraction, decrease the translocation rate or increase nucleotide residence time are generally available. DNA polymerases are sometimes classified into six main groups based upon various phylogenetic relationships, e.g., with *E. coli* Pol I (class A), *E. coli* Pol II (class B), *E. coli* Pol III (class C), Euryarchaeotic Pol II (class D), human Pol beta (class X), and *E. coli* UmuC/DinB and eukaryotic RAD30/xeroderma pigmentosum variant (class Y). For a review of recent nomenclature, see, e.g., Burgers et al. (2001) "Eukaryotic DNA polymerases: proposal for a revised nomenclature" J Biol Chem. 276(47):43487-90. For a review of polymerases, see, e.g., Hübscher et al. (2002) "Eukaryotic DNA Polymerases" Annual Review of Biochemistry Vol. 71: 133-163; Alba (2001) "Protein Family Review: Replicative DNA Polymerases" Genome Biology 2(1):reviews 3002.1-3002.4; and Steitz (1999) "DNA polymerases: structural diversity and common mechanisms" J Biol Chem 274:17395-17398. The basic mechanisms of action for many polymerases have been determined. The sequences of literally hundreds of polymerases are publicly available, and the crystal structures for many of these have been determined, or can be inferred based upon similarity to solved crystal structures for homologous polymerases. For example, the crystal structure of Φ29, a preferred type of parental enzyme to be modified according to the invention, is available.

[0181] DNA polymerases and their properties are described in detail in, among other places, DNA Replication 2nd Edition, Komberg and Baker, W.H. Freeman, New York, N.Y.

(1991). Known conventional DNA polymerases useful in the invention include, but are not limited to, Pyrococcus furiosus (Pfu) DNA polymerase (Lundberg et al., 1991, Gene, 108: 1, Stratagene), Pyrococcus woesei (Pwo) DNA polymerase (Hinnisdaels et al., 1996, Biotechniques, 20:186-8, Boehringer Mannheim), Thermus thermophilus (Tth) DNA polymerase (Myers and Gelfand 1991, Biochemistry 30:7661), Bacillus stearothermophilus DNA polymerase (Stenesh and McGowan, 1977, Biochim Biophys Acta 475:32), Thermococcus litoralis (Tli) DNA polymerase (also referred to as Vent.TM. DNA polymerase, Cariello et al., 1991, Polynucleotides Res, 19: 4193, New England Biolabs), 9.degree. Nm.TM. DNA polymerase (New England Biolabs), Stoffel fragment, ThermoSequenase.RTM. (Amersham Pharmacia Biotech UK), Therminator.TM. (New England Biolabs), Thermotoga maritima (Tma) DNA polymerase (Diaz and Sabino, 1998 Braz J Med. Res, 31:1239), Thermus aquaticus (Taq) DNA polymerase (Chien et al., 1976, J. Bacteoriol, 127: 1550), DNA polymerase, Pyrococcus kodakaraensis KOD DNA polymerase (Takagi et al., 1997, Appl. Environ. Microbiol. 63:4504), JDF-3 DNA polymerase (from thermococcus sp. JDF-3, Patent application WO 0132887), Pyrococcus GB-D (PGB-D) DNA polymerase (also referred as Deep Vent.TM. DNA polymerase, Juncosa-Ginesta et al., 1994, Biotechniques, 16:820, New England Biolabs), UlTma DNA polymerase (from thermophile Thermotoga maritima; Diaz and Sabino, 1998 Braz J. Med. Res, 31:1239; PE Applied Biosystems), Tgo DNA polymerase (from thermococcus gorgonarius, Roche Molecular Biochemicals), E. coli DNA polymerase I (Lecomte and Doubleday, 1983, Polynucleotides Res. 11:7505), T7 DNA polymerase (Nordstrom et al., 1981, J Biol. Chem. 256:3112), and archaeal DP11/DP2 DNA polymerase II (Cann et al., 1998, Proc Natl Acad. Sci. USA 95:14250-5).

[0182] Reverse transcriptases useful in the invention include, but are not limited to, reverse transcriptases from HIV, HTLV-1, HTLV-II, FeLV, FIV, SIV, AMV, MMTV, MoMuLV and other retroviruses (see Levin, Cell 88:5-8 (1997); Verma, Biochim Biophys Acta. 473:1-38 (1977); Wu et al., CRC Crit Rev Biochem. 3:289-347 (1975)).

[0183] In preferred embodiments, the polymerases employed during the sequencing processes, and optionally during pre-sequencing synthesis, will typically possess strand-displacement activity to displace any primers downstream of the primer at which the strand synthesis is initiated. 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. 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, 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. Similarly, polymerases having enhanced activity for labeled nucleotides are also desirable. Examples of polymerase enzymes for use in various aspects of the invention include, e.g., those described in U.S. Patent Application Nos. 11/645,125, filed December 21, 2006; 11/645,135, filed December 21, 2006; 12/384,112, filed March 30, 2009; 61/094,843, filed September 5, 2008; and 61/072,645, filed March 31, 2008; as well as U.S. Patent Publication No. 20070196846 (the full disclosures of which are incorporated herein by reference in their entireties for all purposes), such as the E375Y/K512Y/T368F mutant polymerase described in the foregoing.

[0184] In addition to wild-type polymerases, chimeric polymerases made from a mosaic of different sources can be used. For example, Φ29 polymerases made taking sequences from more than one parental polymerase into account can be used as a starting point for mutating the polymerases of the invention. This can done using consideration of similarity regions between the polymerases to define consensus sequences that are used in the chimera, or can be done using gene shuffling technologies in which multiple Φ29-related polymerases are randomly or semi-randomly shuffled via available gene shuffling techniques (e.g., via "family gene shuffling"; see Crameri et al. (1998) "DNA shuffling of a family of genes from diverse species accelerates directed evolution" Nature 391:288–291; Clackson et al. (1991) "Making antibody fragments using phage display libraries" Nature 352:624-628; Gibbs et al. (2001) "Degenerate oligonucleotide gene shuffling (DOGS): a method for enhancing the frequency of

recombination with family shuffling" Gene 271:13-20; and Hiraga and Arnold (2003) "General method for sequence-independent site-directed chimeragenesis: J. Mol. Biol. 330:287-296). In these methods, the recombination points can be predetermined such that the gene fragments assemble in the correct order. However, the combinations, e.g., chimeras, can be formed at random. For example, using methods described in Clarkson, et al., 5 gene chimeras, e.g., comprising segments of a Phi29 polymerase, a PZA polymerase, a M2 polymerase, a B103 polymerase, and a GA-1 polymerase, with improved branching fractions can be generated.

[0185] Available DNA polymerase enzymes have also been modified in any of a variety of ways, e.g., to simplify production by making protease digested enzyme fragments such as the Klenow fragment recombinant, etc. As noted, polymerases have also been modified to confer improvements in specificity, processivity, and improved retention time of labeled nucleotides in polymerase-DNA-nucleotide complexes (e.g., 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.) and to improve surface-immobilized enzyme activities (e.g., 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.). Any of these available polymerases can be modified in accordance with the invention to increase the average branching fraction, delay translocation and/or increase nucleotide or nucleotide analog residence time.

[0186] The polymerase mutations and mutational strategies noted herein can be combined with available mutations and mutational strategies to confer additional improvements in, e.g., nucleotide analog specificity, enzyme processivity and the like. For example, the mutations and mutational strategies herein can be combined with those taught in, e.g., 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. This combination of mutations/ mutational strategies can be used to impart several simultaneous improvements to a polymerase (decreased branch fraction formation, improved specificity, improved processivity, improved

retention time, improved stability of the closed complex, etc.). In addition, polymerases can be further modified for application-specific reasons, such as to improve activity of the enzyme when bound to a surface, as taught, e.g., 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., or to include purification or handling tags as is taught in the cited references and as is common in the art.

[0187] Specific mutations noted herein can be used alone or in combination with each other and/or with available mutations as described in the references noted above, or can be used in polymerases that lack such previously described mutations.

[0188] Many such polymerases that are suitable for modification are available, e.g., for use in sequencing, labeling and amplification technologies. For example, Human DNA Polymerase Beta is available from R&D systems. DNA polymerase I is available from Epicenter, GE Health Care, Invitrogen, New England Biolabs, Promega, Roche Applied Science, Sigma Aldrich and many others. The Klenow fragment of DNA Polymerase I is available in both recombinant and protease digested versions, from, e.g., Ambion, Chimerx, eEnzyme LLC, GE Health Care, Invitrogen, New England Biolabs, Promega, Roche Applied Science, Sigma Aldrich and many others. Φ29 DNA polymerase is available from e.g., Epicentre. Poly A polymerase, reverse transcriptase, Sequenase, SP6 DNA polymerase, T4 DNA polymerase, T7 DNA polymerase, and a variety of thermostable DNA polymerases (Taq, hot start, titanium Taq, etc.) are available from a variety of these and other sources. Recent commercial DNA polymerases include Phusion™ High-Fidelity DNA Polymerase, available from New England Biolabs; GoTaq® Flexi DNA Polymerase, available from Promega; RepliPHI™ Φ29 DNA Polymerase, available from Epicentre Biotechnologies; PfuUltra™ Hotstart DNA Polymerase, available from Stratagene; KOD HiFi DNA Polymerase, available from Novagen; and many others. Biocompare(dot)com provides comparisons of many different commercially available polymerases.

[0189] DNA polymerases that are preferred substrates for mutation to increase branch rates and/or decrease translocation rate include Taq polymerases, *E. coli* DNA Polymerase 1,

Klenow fragment, reverse transcriptases, Φ29 related polymerases including wild type Φ29 polymerase and derivatives of such polymerases such as exonuclease altered forms, T7 DNA polymerase, T5 DNA polymerase, an RB69 polymerase, etc.

[0190] In one aspect, the modified recombinant polymerases that exhibit increased branching fractions, delayed translocation and/or increased nucleotide or nucleotide analog residence time are Φ29-type DNA polymerases. For example, the modified recombinant DNA polymerases can be homologous to a wild-type Φ29 DNA polymerase, e.g., as described in U.S. Patent Nos. 5,001,050, 5,198,543, or 5,576,204. Alternately, the modified recombinant DNA polymerase can be homologous to other Φ29-type DNA polymerases, such as B103, GA-1, PZA, Φ15, BS32, M2Y, Nf, G1, Cp-1, PRD1, PZE, SF5, Cp-5, Cp-7, PR4, PR5, PR722, L17, Φ21, or the like. For nomenclature, see also, Meijer et al. (2001) "Φ29 Family of Phages" Microbiology and Molecular Biology Reviews, 65(2):261-287.

[0191] As discussed, various polymerases of the invention can incorporate one or more nucleotide analogs into a growing oligonucleotide chain. Upon incorporation, the analog can leave a residue that is the same as or different than a natural nucleotide in the growing oligonucleotide (the polymerase can incorporate any non-standard moiety of the analog, or can cleave it off during incorporation into the oligonucleotide). A "nucleotide analog" herein is a compound, that, in a particular application, functions in a manner similar or analogous to a naturally occurring nucleoside triphosphate (a "nucleotide"), and does not otherwise denote any particular structure. A nucleotide analog is an analog other than a standard naturally occurring nucleotide, i.e., other than A, G, C, T, or U, though upon incorporation into the oligonucleotide, the resulting residue in the oligonucleotide can be the same as (or different from) an A, G, C, T, or U residue.

[0192] Nucleotide analogs can also be modified to achieve any of the improved properties desired. For example, various linkers or other substituents can be incorporated into analogs that have the effect of altering the branching rate, residence time or improving processivity. Modifications to the analogs can include extending the phosphate chains, e.g., to include a hexa- or heptaphosphate group, and/or adding chemical linkers to extend the distance between the nucleotide base and the dye molecule, e.g., fluorescent dye molecule. As

described in detail herein, modifications to the analogs can include altering the analog such that the analog is unincorporatable. For example, nucleotide analogs of the invention can possess unhydrolyzable groups within the phosphate chain, such that the phosphoester linkage between the analog and the primer strand cannot be formed.

[0193] Many nucleotide analogs are available and can be incorporated by the polymerases of the invention. These include analog structures with core similarity to naturally occurring nucleotides, such as those that comprise one or more substituents on a phosphate, sugar or base moiety of the nucleoside or nucleotide relative to a naturally occurring nucleoside or nucleotide. In one embodiment, the nucleotide analog includes three phosphate containing groups; for example, the analog can be a labeled nucleoside triphosphate analog and/or an α-thiophosphate nucleotide analog having three phosphate groups. In one embodiment, a nucleotide analog can include one or more extra phosphate containing groups, relative to a nucleoside triphosphate. For example, a variety of nucleotide analogs that comprise, e.g., from 4-6 or more phosphates are described in detail in U.S. Patent Application No. 11/241,809, filed September 29, 2005, and incorporated herein by reference in its entirety for all purposes. Other exemplary useful analogs, including tetraphosphate and pentaphosphate analogs, are described in U.S. Patent 7,041,812, incorporated herein by reference in its entirety for all purposes.

[0194] Nucleotide analogs of the present invention may comprise any of a variety of detectable labels. Detectable labels generally denote a chemical moiety that provides a basis for detection of the analog compound separate and apart from the same compound lacking such a labeling group. Examples of labels include, e.g., optical labels, e.g., labels that impart a detectable optical property to the analog, electrochemical labels, e.g., labels that impart a detectable electrical or electrochemical property to the analog, and physical labels, e.g., labels that impart a different physical or spatial property to the analog, e.g., a mass tag or molecular volume tag. In some cases individual labels or combinations may be used that impart more than one of the aforementioned properties to the analogs of the invention.

[0195] Optionally, the labeling groups incorporated into the analogs comprise optically detectable moieties, such as luminescent, chemiluminescent, fluorescent, fluorogenic, chromophoric and/or chromogenic moieties, with fluorescent and/or fluorogenic labels being

preferred. A variety of different label moieties are readily employed in nucleotide analogs. Such groups include fluorescein labels, rhodamine labels, cyanine labels (i.e., Cy3, Cy5, and the like, generally available from the Amersham Biosciences division of GE Healthcare), the Alexa family of fluorescent dyes and other fluorescent and fluorogenic dyes available from Molecular Probes/Invitrogen, Inc. and described in 'The Handbook — A Guide to Fluorescent Probes and Labeling Technologies, Tenth Edition' (2005) (available from Invitrogen, Inc./Molecular Probes). A variety of other fluorescent and fluorogenic labels for use with nucleoside polyphosphates, and which would be applicable to the nucleotide analogs incorporated by the polymerases of the present invention, are described in, e.g., U.S. Patent Application Publication No. 2003/0124576, previously incorporated herein by reference in its entirety for all purposes.

[0196] Additional details regarding analogs and methods of making such analogs can be found in U.S. Patent Application No. 11/241,809, filed September 29, 2005, and incorporated herein by reference in its entirety for all purposes.

[0197] Thus, in one illustrative example, the analog can be a phosphate analog (e.g., an analog that has more than the typical number of phosphates found in nucleoside triphosphates) that include, e.g., an Alexa dye label. For example, an Alexa488 dye can be labeled on a delta phosphate of a tetraphosphate analog (denoted, e.g., A488dC4P or A488dA4P, for the Alexa488 labeled tetraphosphate analogs of C and A, respectively), or an Alexa568 or Alexa633 dye can be used (e.g., A568dC4P and A633dC4P, respectively, for labeled tetraphosphate analogs of C or A568dT6P for a labeled tetraphosphate analog of T), or an Alexa546 dye can be used (e.g., A546dG4P), or an Alexa594 dye can be used (e.g., A594dT4P). An Alexa555 dye (A555dC6P), an Alexa 647 dye (A647d6GP) and/or an Alexa660 dye (A660dA6P) can be used in, e.g., single molecule sequencing. Similarly, to facilitate color separation, a pair of fluorophores exhibiting FRET (fluorescence resonance energy transfer) can be labeled on a delta phosphate of a tetraphosphate analog (denoted, e.g., FAM-amb-A532dG4P or FAM-amb-A594dT4P).

[0198] Polymerases of the invention that exhibit branching phenotypes will display a branching fraction that is particular to the nucleotide analog included in the polymerization

reaction. For example, a polymerase of the invention may exhibit different branching fractions for A488dC4P than A488dC6P due, e.g., to the size differences between 4P and 6P analogs. It will be appreciated that polymerases of the present invention can be modified such that it exhibits the desired branching fraction for a particular nucleotide analog, e.g., a dye-labeled nucleotide analog with a particular number of phosphate groups.

APPLICATIONS FOR ENHANCED NUCLEIC ACID SEQUENCING

[0199] Polymerases of the invention, e.g., modified recombinant polymerases, are used in combination with nucleotides and/or nucleotide analogs, and nucleic acid templates (DNA or RNA) to copy the template nucleic acid. That is, a mixture of the polymerase, nucleotides/analogs, and optionally and other appropriate reagents, the template and a replication initiating moiety (e.g., primer) is reacted such that the polymerase synthesizes nucleic acid (e.g., extends the primer) in a template-dependent manner. The replication initiating moiety can be a standard oligonucleotide primer, or, alternatively, a component of the template, e.g., the template can be a self-priming single stranded DNA, a nicked double stranded DNA, or the like. Similarly, a terminal protein can serve as a initiating moiety. At least one nucleotide analog can be incorporated into the DNA. The template DNA can be a linear or circular DNA, and in certain applications, is desirably a circular template (e.g., for rolling circle replication or for sequencing of circular templates). Optionally, the composition can be present in an automated DNA replication and/or sequencing system.

[0200] Incorporation of labeled nucleotide analogs by the polymerases of the invention is particularly useful in a variety of different nucleic acid analyses, including real-time monitoring of DNA polymerization. The label can itself be incorporated, or more preferably, can be released during incorporation of the analog. For example, analog incorporation can be monitored in real-time by monitoring label release during incorporation of the analog by the polymerase. The portion of the analog that is incorporated can be the same as a natural nucleotide, or can include features of the analog that differ from a natural nucleotide.

[0201] In general, label incorporation or release can be used to indicate the presence and composition of a growing nucleic acid strand, e.g., providing evidence of template replication/amplification and/or sequence of the template. Signaling from the incorporation can

be the result of detecting labeling groups that are liberated from the incorporated analog, e.g., in a solid phase assay, or can arise upon the incorporation reaction. For example, in the case of FRET labels where a bound label is quenched and a free label is not, release of a label group from the incorporated analog can give rise to a fluorescent signal. Alternatively, the enzyme may be labeled with one member of a FRET pair proximal to the active site, and incorporation of an analog bearing the other member will allow energy transfer upon incorporation. The use of enzyme bound FRET components in nucleic acid sequencing applications is described, e.g., in U.S. Patent Application Publication No. 2003/0044781, incorporated herein by reference.

[0202] In one example reaction of interest, a polymerase reaction can be isolated within an extremely small observation volume that effectively results in observation of individual polymerase molecules. In a preferred aspect, such small observation volumes are provided by immobilizing the polymerase enzyme within a structural confinement, such as a nanohole (an aperture of less than 1 μM diameter through which a synthesis complex can be illuminated by optical energy) (See, e.g., co-pending Published U.S. Patent Application No. 2007-0188750, and published International Patent Application No. WO 2007/095119, the full disclosures of which are incorporated herein by reference in their entirety for all purposes) or nanoholes that additionally provide optical confinement, such as a zero-mode waveguide (ZMW). For a description of ZMWs and their application in single molecule analyses, and particularly nucleic acid sequencing, see, e.g., U.S. Patent Application Publication No. 2003/0044781, and U.S. Patent No. 6,917,726, each of which is incorporated herein by reference in its entirety for all purposes. See also Levene et al. (2003) "Zero-mode waveguides for single-molecule analysis at high concentrations" Science 299:682-686 and U.S. Patent Nos. 7,056,676, 7,056,661, 7,052,847, and 7,033,764, the full disclosures of which are incorporated herein by reference in their entirety for all purposes.

[0203] In general, a polymerase enzyme is complexed with the template strand in the presence of one or more nucleotides and/or one or more nucleotide analogs. For example, in certain embodiments, labeled analogs are present representing analogous compounds to each of the four natural nucleotides, A, T, G and C, e.g., in separate polymerase reactions, as in classical Sanger sequencing, or multiplexed together, e.g., in a single reaction, as in multiplexed sequencing approaches. When a particular base in the template strand is

encountered by the polymerase during the polymerization reaction, it complexes with an available analog that is complementary to such nucleotide, and incorporates that analog into the nascent and growing nucleic acid strand. In one aspect, incorporation can result in a label being released, e.g., in polyphosphate analogs, cleaving between the α and β phosphorus atoms in the analog, and consequently releasing the labeling group (or a portion thereof). The incorporation event is detected, either by virtue of a longer presence of the analog and, thus, the label, in the complex, or by virtue of release of the label group into the surrounding medium. Where different labeling groups are used for each of the types of analogs, e.g., A, T, G or C, identification of a label of an incorporated analog allows identification of that analog and consequently, determination of the complementary nucleotide in the template strand being processed at that time. Sequential reaction and monitoring permits a real-time monitoring of the polymerization reaction and determination of the sequence of the template nucleic acid. As noted above, in particularly preferred aspects, the polymerase enzyme/template complex is provided immobilized within an optical confinement that permits observation of an individual complex, e.g., a nanohole or zero mode waveguide.

[0204] In addition to their use in sequencing, the polymerases and/or reaction conditions of the invention are also useful in a variety of other genotyping analyses, e.g., SNP genotyping using single base extension methods, real time monitoring of amplification, e.g., RT-PCR methods, and the like. Further details regarding sequencing and nucleic acid amplification can be found, e.g., in Sambrook et al., Molecular Cloning - A Laboratory Manual (3rd Ed.), Vol. 1-3, Cold Spring Harbor Laboratory, Cold Spring Harbor, New York, 2000; Current Protocols in Molecular Biology, F.M. Ausubel et al., eds., Current Protocols, a joint venture between Greene Publishing Associates, Inc. and John Wiley & Sons, Inc., (supplemented through 2006); and PCR Protocols: A Guide to Methods and Applications (Innis *et al.* eds) Academic Press Inc. San Diego, CA (1990).

MOLECULAR MODELING-BASED MODIFICATION OF POLYMERASES TO INCREASE BRANCHING FRACTION, DELAY TRANSLOCATION OR INCREASE NUCLEOTIDE ANALOG RETENTION TIME

Structure-based design of recombinant polymerases

[0205] Structural data for a polymerase can be used to conveniently identify amino acid residues as candidates for mutagenesis to create recombinant polymerases, e.g., having modified active site regions that increase the branching fractions. For example, analysis of the three dimensional structure of a polymerase such as Φ29 can identify residues that are particularly relevant to branching, translocation and/or nucleotide residence time properties of the polymerase.

[0206] The three-dimensional structures of a large number of DNA polymerases have been determined by x-ray crystallography and nuclear magnetic resonance (NMR) spectroscopy, including the structures of polymerases with bound templates, nucleotides, and/or nucleotide analogs. Many such structures are freely available for download from the Protein Data Bank, at (www(dot)rcsb(dot)org/pdb. Structures, along with domain and homology information, are also freely available for search and download from the National Center for Biotechnology Information's Molecular Modeling DataBase, at www(dot)ncbi(dot)nlm(dot)nih(dot)gov/Structure/MMDB/mmdb(dot)shtml. The structures of additional polymerases can be modeled, for example, based on homology of the polymerases with polymerases whose structures have already been determined. Alternatively, the structure of a given polymerase, optionally complexed with a template and/or nucleotide analog, or the like, can be determined.

[0207] Techniques for crystal structure determination are well known. See, for example, McPherson (1999) Crystallization of Biological Macromolecules Cold Spring Harbor Laboratory; Bergfors (1999) Protein Crystallization International University Line; Mullin (1993) Crystallization Butterwoth-Heinemann; Stout and Jensen (1989) X-ray structure determination: a practical guide, 2nd Edition Wiley Publishers, New York; Ladd and Palmer (1993) Structure determination by X-ray crystallography, 3rd Edition Plenum Press, NewYork; Blundell and Johnson (1976) Protein Crystallography Academic Press, New York; Glusker and Trueblood (1985) Crystal structure analysis: A primer, 2nd Ed. Oxford University Press, NewYork; International Tables for Crystallography, Vol. F, Crystallography of Biological Macromolecules; McPherson (2002) Introduction to Macromolecular Crystallography Wiley-Liss; McRee and David (1999) Practical Protein Crystallography, Second Edition Academic Press; Drenth (1999) Principles of Protein X-Ray Crystallography (Springer Advanced Texts in

Chemistry) Springer-Verlag; Fanchon and Hendrickson (1991) Chapter 15 of Crystallographic Computing, Volume 5 IUCr/Oxford University Press; Murthy (1996) Chapter 5 of Crystallographic Methods and Protocols Humana Press; Dauter et al. (2000) "Novel approach to phasing proteins: derivatization by short cryo-soaking with halides" Acta Cryst.D56:232-237; Dauter (2002) "New approaches to high-throughput phasing" Curr. Opin. Structural Biol. 12:674–678; Chen et al. (1991) "Crystal structure of a bovine neurophysin-II dipeptide complex at 2.8 Å determined from the single-wavelength anomalous scattering signal of an incorporated iodine atom" Proc. Natl Acad. Sci. USA, 88:4240–4244; and Gavira et al. (2002) "*Ab initio* crystallographic structure determination of insulin from protein to electron density without crystal handling" Acta Cryst.D58:1147-1154.

[0208] In addition, a variety of programs to facilitate data collection, phase determination, model building and refinement, and the like are publicly available. Examples include, but are not limited to, the HKL2000 package (Otwinowski and Minor (1997) "Processing of X-ray Diffraction Data Collected in Oscillation Mode" Methods in Enzymology 276:307-326), the CCP4 package (Collaborative Computational Project (1994) "The CCP4 suite: programs for protein crystallography" Acta Crystallogr D 50:760-763), SOLVE and RESOLVE (Terwilliger and Berendzen (1999) Acta Crystallogr D 55 (Pt 4):849-861), SHELXS and SHELXD (Schneider and Sheldrick (2002) "Substructure solution with SHELXD" Acta Crystallogr D Biol Crystallogr 58:1772-1779), Refmac5 (Murshudov et al. (1997) "Refinement of Macromolecular Structures by the Maximum-Likelihood Method" Acta Crystallogr D 53:240-255), PRODRG (van Aalten et al. (1996) "PRODRG, a program for generating molecular topologies and unique molecular descriptors from coordinates of small molecules" J Comput Aided Mol Des 10:255-262), and O (Jones et al. (1991) "Improved methods for building protein models in electron density maps and the location of errors in these models" Acta Crystallogr A 47 (Pt 2):110-119).

[0209] Techniques for structure determination by NMR spectroscopy are similarly well described in the literature. See, e.g., Cavanagh et al. (1995) Protein NMR Spectroscopy: Principles and Practice, Academic Press; Levitt (2001) Spin Dynamics: Basics of Nuclear Magnetic Resonance, John Wiley & Sons; Evans (1995) Biomolecular NMR Spectroscopy, Oxford University Press; Wüthrich (1986) NMR of Proteins and Nucleic Acids (Baker Lecture

Series), Kurt Wiley-Interscience; Neuhaus and Williamson (2000) The Nuclear Overhauser Effect in Structural and Conformational Analysis, 2nd Edition, Wiley-VCH; Macomber (1998) A Complete Introduction to Modern NMR Spectroscopy, Wiley-Interscience; Downing (2004) Protein NMR Techniques (Methods in Molecular Biology), 2nd edition, Humana Press; Clore and Gronenborn (1994) NMR of Proteins (Topics in Molecular and Structural Biology), CRC Press; Reid (1997) Protein NMR Techniques, Humana Press; Krishna and Berliner (2003) Protein NMR for the Millenium (Biological Magnetic Resonance), Kluwer Academic Publishers; Kiihne and De Groot (2001) Perspectives on Solid State NMR in Biology (Focus on Structural Biology, 1), Kluwer Academic Publishers; Jones et al. (1993) Spectroscopic Methods and Analyses: NMR, Mass Spectrometry, and Related Techniques (Methods in Molecular Biology, Vol. 17), Humana Press; Goto and Kay (2000) Curr. Opin. Struct. Biol. 10:585; Gardner (1998) Annu. Rev. Biophys. Biomol. Struct. 27:357; Wüthrich (2003) Angew. Chem. Int. Ed. 42:3340; Bax (1994) Curr. Opin. Struct. Biol. 4:738; Pervushin et al. (1997) Proc. Natl. Acad. Sci. U.S.A. 94:12366; Fiaux et al. (2002) Nature 418:207; Fernandez and Wider (2003) Curr. Opin. Struct. Biol. 13:570; Ellman et al. (1992) J. Am. Chem. Soc. 114:7959; Wider (2000) BioTechniques 29:1278-1294; Pellecchia et al. (2002) Nature Rev. Drug Discov. (2002) 1:211-219; Arora and Tamm (2001) Curr. Opin. Struct. Biol. 11:540-547; Flaux et al. (2002) Nature 418:207-211; Pellecchia et al. (2001) J. Am. Chem. Soc. 123:4633-4634; and Pervushin et al. (1997) Proc. Natl. Acad. Sci. USA 94:12366-12371.

[0210] The structure of a polymerase, or polymerase bound to a DNA or with a given nucleotide analog incorporated into the active site can, as noted, be directly determined, e.g., by x-ray crystallography or NMR spectroscopy, or the structure can be modeled based on the structure of the polymerase and/or a structure of a polymerase with a natural nucleotide bound. The active site or other relevant domain of the polymerase can be identified, for example, by homology with other polymerases, examination of polymerase-template or polymerase-nucleotide co-complexes, biochemical analysis of mutant polymerases, and/or the like. The position of a nucleotide analog (as opposed to an available nucleotide structure) in the active site can be modeled, for example, by projecting the location of non-natural features of the analog (e.g., additional phosphate or phosphonate groups in the phosphorus containing chain linked to the nucleotide, e.g., tetra, penta or hexa phosphate groups, detectable labeling groups,

e.g., fluorescent dyes, or the like) based on the previously determined location of another nucleotide or nucleotide analog in the active site.

[0211] Such modeling of the nucleotide analog or template (or both) in the active site can involve simple visual inspection of a model of the polymerase, for example, using molecular graphics software such as the PyMOL viewer (open source, freely available on the World Wide Web at www(dot)pymol(dot)org) or Insight II (commercially available from Accelrys at (www (dot) accelrys (dot) com/products/insight). Alternatively, modeling of the active site complex of the polymerase or a putative mutant polymerase, for example, can involve computer-assisted docking, molecular dynamics, free energy minimization, and/or like calculations. Such modeling techniques have been well described in the literature; see, e.g., Babine and Abdel-Meguid (eds.) (2004) Protein Crystallography in Drug Design, Wiley-VCH, Weinheim; Lyne (2002) "Structure-based virtual screening: An overview" Drug Discov. Today 7:1047-1055; Molecular Modeling for Beginners, at (www (dot) usm (dot) maine (dot) edu/~rhodes/SPVTut/index (dot) html; and Methods for Protein Simulations and Drug Design at (www (dot) dddc (dot) ac (dot) cn/embo04; and references therein. Software to facilitate such modeling is widely available, for example, the CHARMm simulation package, available academically from Harvard University or commercially from Accelrys (at www (dot) accelrys (dot) com), the Discover simulation package (included in Insight II, *supra*), and Dynama (available at (www(dot) cs (dot) gsu (dot) edu/~cscrwh/progs/progs (dot) html). See also an extensive list of modeling software at (www (dot) netsci (dot) org/Resources/Software/Modeling/MMMD/top (dot) html.

[0212] Visual inspection and/or computational analysis of a polymerase model can identify relevant features of the active site or other domain, including, for example, amino acid residues of domains that are in close proximity to one another (to stabilize inter-domain interactions) residues in the active site that interact with the nucleotide or analog, or that modulate how large a binding pocket for the analog is relative to the analog. That is, inter-domain amino acid contacts can stabilize the closed complex, and/or the size or composition (e.g., position of charged or hydrophobic residues) of the binding pocket in the active site can control entry and release of the nucleotide, which can affect branching rate. A residue can, for example, be deleted or replaced with a residue having a different (smaller, larger, ionic, non-

ionic, etc.) side chain. Similarly, residues that can be altered to introduce desirable interactions with the nucleotide analog can be identified to reduce branching. Such a residue can be replaced with a residue that is complementary with, e.g., a non-natural feature of the analog, for example, with a residue that can hydrogen bond to the analog (e.g., serine, threonine, histidine, asparagine, or glutamine), a hydrophobic residue that can interact with a hydrophobic group on the analog, an aromatic residue that can provide favorable hydrophobic interactions with a group on the analog (e.g., a fluorophore), an aromatic residue that can engage in a π-π or edge-face stacking interaction with an aromatic group in the analog, a residue that can engage in a cation-π interaction with the analog, or a charged residue (e.g., aspartic or glutamic acid, or lysine, arginine, or histidine) that can electrostatically interact with an oppositely charged moiety on the analog (e.g., an additional phosphate group).

[0213] Thus, in addition to methods of using the polymerases and other compositions herein, the present invention also includes methods of making the polymerases. As described, methods of making a recombinant DNA polymerase can include structurally modeling a first polymerase, e.g., using any available crystal structure and molecular modeling software or system. Based on the modeling, one or more feature affecting closed complex stability, or nucleotide access or removal to or from the active site (and, thereby, branching) and/or binding of a DNA or nucleotide analog within the active site region is identified. These residues can be, e.g., in the active site, an exonuclease, TPR2 or thumb domain (or interface between domains) or proximal to such domains. The DNA polymerase is mutated to include non-natural residues at such positions, and then screened for an activity of interest.

Mutating Polymerases

[0214] Various types of mutagenesis are optionally used in the present invention, e.g., to modify polymerases to produce variants, e.g., in accordance with polymerase models and model predictions as discussed above, or using random or semi-random mutational approaches. In general, any available mutagenesis procedure can be used for making polymerase mutants. Such mutagenesis procedures optionally include selection of mutant nucleic acids and polypeptides for one or more activity of interest (e.g., decreased branch fraction, increased or decreased complex stability, improved processivity, and/or improved k_{off}, K_m, V_{max}, k_{cat} etc.,

e.g., for a given nucleotide analog). Procedures that can be used include, but are not limited to: site-directed point mutagenesis, random point mutagenesis, in vitro or in vivo homologous recombination (DNA shuffling and combinatorial overlap PCR), mutagenesis using uracil containing templates, oligonucleotide-directed mutagenesis, phosphorothioate-modified DNA mutagenesis, mutagenesis using gapped duplex DNA, point mismatch repair, mutagenesis using repair-deficient host strains, restriction-selection and restriction-purification, deletion mutagenesis, mutagenesis by total gene synthesis, degenerate PCR, double-strand break repair, and many others known to persons of skill. The starting polymerase for mutation can be any of those noted herein, including available polymerase mutants such as those identified e.g., in WO 2007/076057 POLYMERASES FOR NUCLEOTIDE ANALOG INCORPORATION by Hanzel 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.

[0215] Optionally, mutagenesis can be guided by known information from a naturally occurring polymerase molecule, or of a known altered or mutated polymerase (e.g., using an existing mutant polymerase as noted in the preceding references), e.g., sequence, sequence comparisons, physical properties, crystal structure and/or the like as discussed above. However, in another class of embodiments, modification can be essentially random (e.g., as in classical or "family" DNA shuffling, see, e.g., Crameri et al. (1998) "DNA shuffling of a family of genes from diverse species accelerates directed evolution" Nature 391:288–291).

[0216] Additional information on mutation formats is found in: Sambrook et al., Molecular Cloning - A Laboratory Manual (3rd Ed.), Vol. 1-3, Cold Spring Harbor Laboratory, Cold Spring Harbor, New York, 2000 ("Sambrook"); Current Protocols in Molecular Biology, F.M. Ausubel et al., eds., Current Protocols, a joint venture between Greene Publishing Associates, Inc. and John Wiley & Sons, Inc., (supplemented through 2006) ("Ausubel")) and PCR Protocols A Guide to Methods and Applications (Innis *et al.* eds) Academic Press Inc. San Diego, CA (1990) (Innis). The following publications and references cited within provide additional detail on mutation formats: Arnold, *Protein engineering for unusual environments*,

Current Opinion in Biotechnology 4:450-455 (1993); Bass et al., *Mutant Trp repressors with new DNA-binding specificities*, Science 242:240-245 (1988); Bordo and Argos (1991) *Suggestions for "safe Residue Substitutions in Site-directed Mutagenesis* 217:721-729; Botstein & Shortle, *Strategies and applications of in vitro mutagenesis*, Science 229:1193-1201(1985); Carter et al., *Improved oligonucleotide site-directed mutagenesis using M13 vectors*, Nucl. Acids Res. 13: 4431-4443 (1985); Carter, *Site-directed mutagenesis*, Biochem. J. 237:1-7 (1986); Carter, *Improved oligonucleotide-directed mutagenesis using M13 vectors*, Methods in Enzymol. 154: 382-403 (1987); Dale et al., *Oligonucleotide-directed random mutagenesis using the phosphorothioate method*, Methods Mol. Biol. 57:369-374 (1996); Eghtedarzadeh & Henikoff, *Use of oligonucleotides to generate large deletions*, Nucl. Acids Res. 14: 5115 (1986); Fritz et al., *Oligonucleotide-directed construction of mutations: a gapped duplex DNA procedure without enzymatic reactions in vitro*, Nucl. Acids Res. 16: 6987-6999 (1988); Grundström et al., *Oligonucleotide-directed mutagenesis by microscale 'shot-gun' gene synthesis*, Nucl. Acids Res. 13: 3305-3316 (1985); Hayes (2002) *Combining Computational and Experimental Screening for rapid Optimization of Protein Properties* PNAS 99(25) 15926-15931; Kunkel, *The efficiency of oligonucleotide directed mutagenesis*, in Nucleic Acids & Molecular Biology (Eckstein, F. and Lilley, D.M.J. eds., Springer Verlag, Berlin)) (1987); Kunkel, *Rapid and efficient site-specific mutagenesis without phenotypic selection*, Proc. Natl. Acad. Sci. USA 82:488-492 (1985); Kunkel et al., *Rapid and efficient site-specific mutagenesis without phenotypic selection*, Methods in Enzymol. 154, 367-382 (1987); Kramer et al., *The gapped duplex DNA approach to oligonucleotide-directed mutation construction*, Nucl. Acids Res. 12: 9441-9456 (1984); Kramer & Fritz *Oligonucleotide-directed construction of mutations via gapped duplex DNA*, Methods in Enzymol. 154:350-367 (1987); Kramer et al., *Point Mismatch Repair*, Cell 38:879-887 (1984); Kramer et al., *Improved enzymatic in vitro reactions in the gapped duplex DNA approach to oligonucleotide-directed construction of mutations*, Nucl. Acids Res. 16: 7207 (1988); Ling et al., *Approaches to DNA mutagenesis: an overview*, Anal Biochem. 254(2): 157-178 (1997); Lorimer and Pastan Nucleic Acids Res. 23, 3067-8 (1995); Mandecki, *Oligonucleotide-directed double-strand break repair in plasmids of Escherichia coli: a method for site-specific mutagenesis*, Proc. Natl. Acad. Sci. USA, 83:7177-7181 (1986); Nakamaye & Eckstein, *Inhibition of restriction endonuclease Nci I cleavage by*

phosphorothioate groups and its application to oligonucleotide-directed mutagenesis, Nucl. Acids Res. 14: 9679-9698 (1986); Nambiar et al., *Total synthesis and cloning of a gene coding for the ribonuclease S protein*, Science 223: 1299-1301 (1984); Sakamar and Khorana, *Total synthesis and expression of a gene for the a-subunit of bovine rod outer segment guanine nucleotide-binding protein (transducin)*, Nucl. Acids Res. 14: 6361-6372 (1988); Sayers et al., *Y-T Exonucleases in phosphorothioate-based oligonucleotide-directed mutagenesis*, Nucl. Acids Res. 16:791-802 (1988); Sayers et al., *Strand specific cleavage of phosphorothioate-containing DNA by reaction with restriction endonucleases in the presence of ethidium bromide*, (1988) Nucl. Acids Res. 16: 803-814; Sieber, et al., Nature Biotechnology, 19:456-460 (2001); Smith, *In vitro mutagenesis*, Ann. Rev. Genet. 19:423-462(1985); Methods in Enzymol. 100: 468-500 (1983); Methods in Enzymol. 154: 329-350 (1987); Stemmer, Nature 370, 389-91 (1994); Taylor et al., *The use of phosphorothioate-modified DNA in restriction enzyme reactions to prepare nicked DNA*, Nucl. Acids Res. 13: 8749-8764 (1985); Taylor et al., *The rapid generation of oligonucleotide-directed mutations at high frequency using phosphorothioate-modified DNA*, Nucl. Acids Res. 13: 8765-8787 (1985); Wells et al., *Importance of hydrogen-bond formation in stabilizing the transition state of subtilisin*, Phil. Trans. R. Soc. Lond. A 317: 415-423 (1986); Wells et al., *Cassette mutagenesis: an efficient method for generation of multiple mutations at defined sites*, Gene 34:315-323 (1985); Zoller & Smith, *Oligonucleotide-directed mutagenesis using M13-derived vectors: an efficient and general procedure for the production of point mutations in any DNA fragment*, Nucleic Acids Res. 10:6487-6500 (1982); Zoller & Smith, *Oligonucleotide-directed mutagenesis of DNA fragments cloned into M13 vectors*, Methods in Enzymol. 100:468-500 (1983); Zoller & Smith, *Oligonucleotide-directed mutagenesis: a simple method using two oligonucleotide primers and a single-stranded DNA template*, Methods in Enzymol. 154:329-350 (1987); Clackson et al. (1991) "Making antibody fragments using phage display libraries" Nature 352:624-628; Gibbs et al. (2001) "Degenerate oligonucleotide gene shuffling (DOGS): a method for enhancing the frequency of recombination with family shuffling" Gene 271:13-20; and Hiraga and Arnold (2003) "General method for sequence-independent site-directed chimeragenesis: J. Mol. Biol. 330:287-296. Additional details on many of the above methods can be found in Methods in

Enzymology Volume 154, which also describes useful controls for trouble-shooting problems with various mutagenesis methods.

Determining Kinetic Parameters

[0217] The polymerases of the invention can be screened or otherwise tested to determine whether the polymerase displays a modified activity for or with a nucleotide analog or template as compared to the first DNA polymerase (e.g., a corresponding wild-type or available mutant polymerase from which the recombinant polymerase of the invention was derived). For example, branch fraction, k_{off}, k_{cat}, K_m, V_{max}, k_{cat}/K_m, V_{max}/K_m, k_{pol}, and/or K_d of the recombinant DNA polymerase for the nucleotide (or analog) or template nucleic acid can be determined. The enzyme perfection metric k_{cat}/K_m is also a useful measure, e.g., for assessing branch rate. k_{cat}/K_m is a measure of substrate binding that leads to product formation (and, thus, includes terms defining binding K_d and inversely predicts branching fraction formation).

[0218] As is well-known in the art, for enzymes obeying simple Michaelis-Menten kinetics, kinetic parameters are readily derived from rates of catalysis measured at different substrate concentrations. The Michaelis-Menten equation, $V=V_{max}[S]([S]+K_m)^{-1}$, relates the concentration of uncombined substrate ([S], approximated by the total substrate concentration), the maximal rate (V_{max}, attained when the enzyme is saturated with substrate), and the Michaelis constant (K_m, equal to the substrate concentration at which the reaction rate is half of its maximal value), to the reaction rate (V). Details regarding K_{off} determination are described above. In general, the dissociation rate can be measured in any manner that detects the polymerase/ DNA complex over time. This includes stopped flow spectroscopy, or even simply by taking aliquots over time and testing for polymerase activity on the template of interest. Free polymerase is captured with a polymerase trap after dissociation, e.g., by incubation in the presence of heparin or an excess of competitor DNA (e.g., non-specific salmon sperm DNA, or the like).

[0219] For many enzymes, K_m is equal to the dissociation constant of the enzyme-substrate complex and is thus a measure of the strength of the enzyme-substrate complex. For such an enzyme, in a comparison of K_ms, a lower K_m represents a complex with stronger binding, while a higher Km represents a complex with weaker binding. The ratio k_{cat}/K_m,

sometimes called the specificity constant, represents the apparent rate constant for combination of substrate with free enzyme. The larger the specificity constant, the more efficient the enzyme is in binding the substrate and converting it to product (this provides an inverse measure of branching rate, as branching rate is the rate at which the enzyme binds substrate (e.g., nucleotide), but does not convert it to product (e.g., a DNA polymer).

[0220] k_{cat} (also called the turnover number of the enzyme) can be determined if the total enzyme concentration ($[E_T]$, i.e., the concentration of active sites) is known, since $V_{max}=k_{cat}[E_T]$. For situations in which the total enzyme concentration is difficult to measure, the ratio V_{max}/K_m is often used instead as a measure of efficiency. K_m and V_{max} can be determined, for example, from a Lineweaver-Burk plot of 1/V against 1/[S], where the y intercept represents $1/V_{max}$, the x intercept $-1/K_m$, and the slope K_m/V_{max}, or from an Eadie-Hofstee plot of V against V/[S], where the y intercept represents V_{max}, the x intercept V_{max}/K_m, and the slope $-K_m$. Software packages such as KinetAsyst™ or Enzfit (Biosoft, Cambridge, UK) can facilitate the determination of kinetic parameters from catalytic rate data.

[0221] For enzymes such as polymerases that have multiple substrates, varying the concentration of only one substrate while holding the others in suitable excess (e.g., effectively constant) concentration typically yields normal Michaelis-Menten kinetics.

[0222] In one embodiment, using pre-steady-state kinetics, the nucleotide concentration dependence of the rate k_{obs} (the observed first-order rate constant for dNTP incorporation) provides an estimate of the K_m for a ground state binding and the maximum rate of polymerization (k_{pol}). The k_{obs} is measured using a burst assay. The results of the assay are fitted with the Burst equation; Product = $A[1-\exp(-k_{obs}*t)]+k_{ss}*t$ where A represents amplitude an estimate of the concentration of the enzyme active site*s, k_{ss} is the observed steady-state rate constant and t is the reaction incubation time. The K_m for dNTP binding to the polymerase-DNA complex and the k_{pol} are calculated by fitting the dNTP concentration dependent change in the k_{obs} using the equation $k_{obs} = (k_{pol}*[S])*(K_m+[S])-1$ where [S] is the substrate concentration. Results are optionally obtained from a rapid-quench experiment (also called a quench-flow measurement), for example, based on the methods described in Johnson (1986) "Rapid kinetic analysis of mechanochemical adenosinetriphosphatases" Methods Enzymol.

134:677-705, Patel et al. (1991) "Pre-steady-state kinetic analysis of processive DNA replication including complete characterization of an exonuclease-deficient mutant" Biochemistry 30(2):511-25, and Tsai and Johnson (2006) "A new paradigm for DNA polymerase specificity" Biochemistry 45(32):9675-87.

[0223] Parameters such as rate of binding of a nucleotide analog or template by the recombinant polymerase, rate of product release by the recombinant polymerase, or branching rate of the recombinant polymerase can also be determined, and optionally compared to that of the first polymerase (e.g., a corresponding wild-type polymerase).

[0224] For a more thorough discussion of enzyme kinetics, see, e.g., Berg, Tymoczko, and Stryer (2002) Biochemistry, Fifth Edition, W. H. Freeman; Creighton (1984) Proteins: Structures and Molecular Principles, W. H. Freeman; and Fersht (1985) Enzyme Structure and Mechanism, Second Edition, W. H. Freeman.

[0225] In one aspect, the improved property of the enzymes of the invention is measured with reference to a model analog or analog set and compared with a given parental enzyme. For example, in the case of enzymes derived from a Φ29 parental enzyme, where the improvement being sought is a decrease in stability of the closed complex, an improved enzyme of the invention (i.e., an enzyme with an increased branching fraction) would have a higher k_{off} than the parental enzyme, e.g., wild type Φ29. While the foregoing may be used as a characterization tool, it in no way is intended as a specifically limiting reaction of the invention.

Screening Polymerases

[0226] Screening or other protocols can be used to determine whether a polymerase displays a modified activity for a nucleotide analog as compared to the first DNA polymerase. For example, k_{off}, k_{cat}, K_m, V_{max}, or k_{cat}/K_m of the recombinant DNA polymerase for the template or nucleotide or analog can be determined as discussed above.

[0227] In one desirable aspect, a library of recombinant DNA polymerases can be made and screened for these properties. For example, a plurality of members of the library can be made to include one or more mutations that increase branching fractions, delay translocation

and/or increase residence time, that are then screened for the properties of interest (e.g., increased branching fraction, delayed translocation or increased residence time). In general, the library can be screened to identify at least one member comprising a modified activity of interest.

[0228] Libraries of polymerases can be either physical or logical in nature. Moreover, any of a wide variety of library formats can be used. For example, polymerases can be fixed to solid surfaces in arrays of proteins. Similarly, liquid phase arrays of polymerases (e.g., in microwell plates) can be constructed for convenient high-throughput fluid manipulations of solutions comprising polymerases. Liquid, emulsion, or gel-phase libraries of cells that express recombinant polymerases can also be constructed, e.g., in microwell plates, or on agar plates. Phage display libraries of polymerases or polymerase domains (e.g., including the active site region or interdomain stability regions) can be produced. Likewise, yeast display libraries can be used. Instructions in making and using libraries can be found, e.g., in Sambrook, Ausubel and Berger, referenced herein.

[0229] For the generation of libraries involving fluid transfer to or from microtiter plates, a fluid handling station is optionally used. Several "off the shelf" fluid handling stations for performing such transfers are commercially available, including e.g., the Zymate systems from Caliper Life Sciences (Hopkinton, MA) and other stations which utilize automatic pipettors, e.g., in conjunction with the robotics for plate movement (e.g., the ORCA® robot, which is used in a variety of laboratory systems available, e.g., from Beckman Coulter, Inc. (Fullerton, CA).

[0230] In an alternate embodiment, fluid handling is performed in microchips, e.g., involving transfer of materials from microwell plates or other wells through microchannels on the chips to destination sites (microchannel regions, wells, chambers or the like). Commercially available microfluidic systems include those from Hewlett-Packard/Agilent Technologies (e.g., the HP2100 bioanalyzer) and the Caliper High Throughput Screening System. The Caliper High Throughput Screening System provides one example interface between standard microwell library formats and Labchip technologies. RainDance Technologies' nanodroplet platform provides another method for handling large numbers of

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

(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.

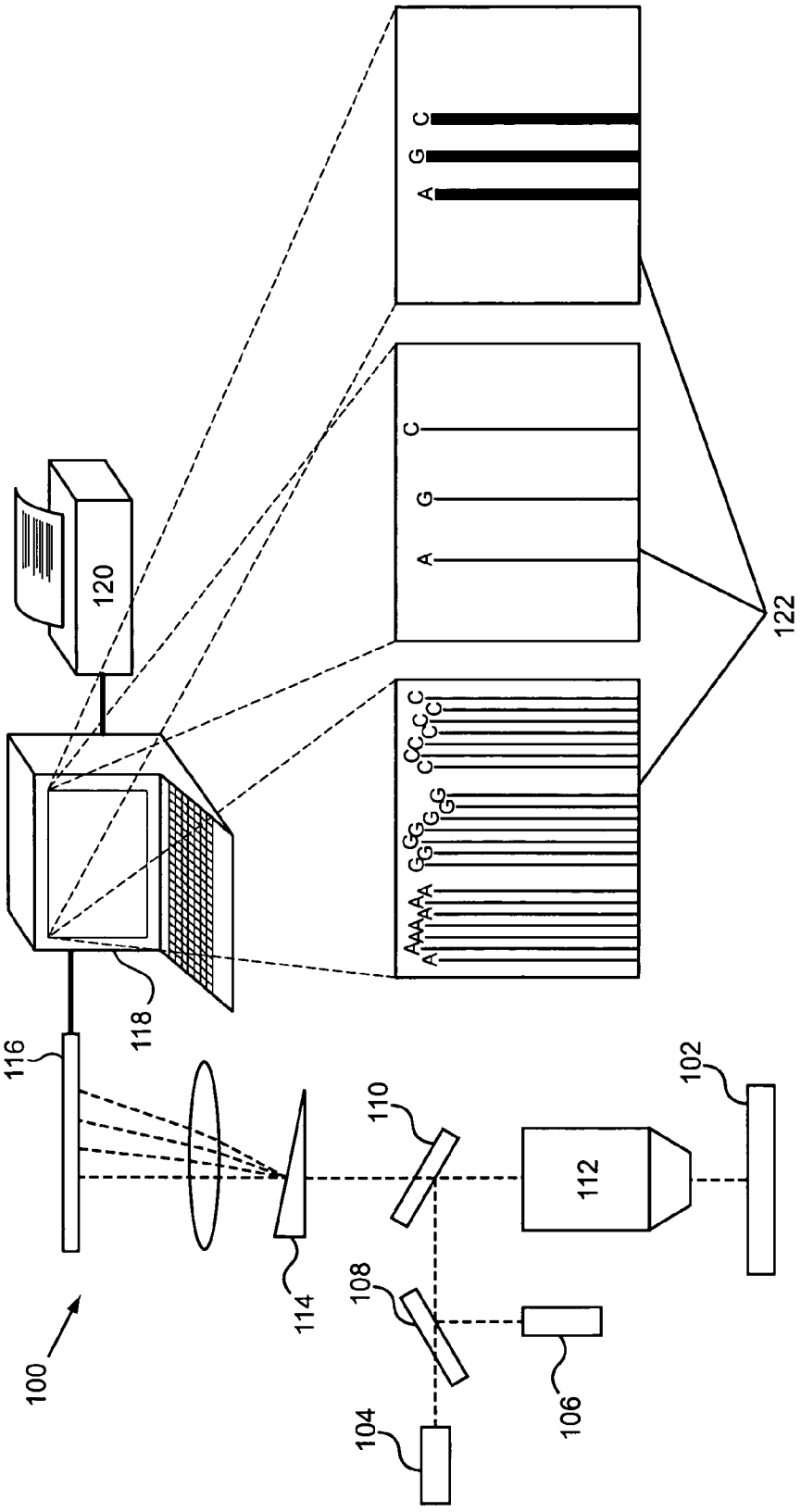

A
G
C
A
G
C
AAAAAAA
GGGGGGG
CCCCCCC
120
122
118
116
102
110
112
114
108
104
106
100

Fig. 1

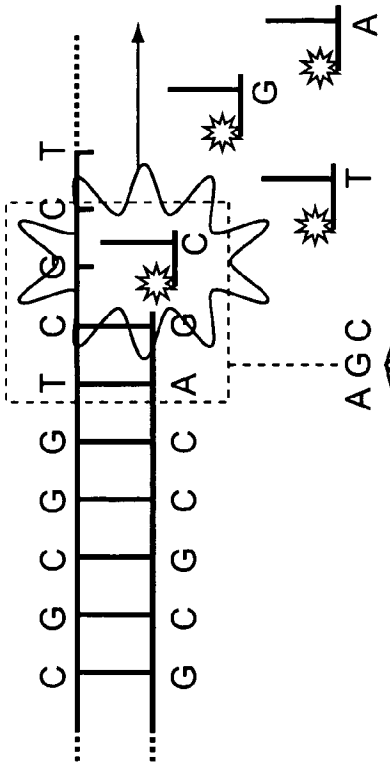

C G C G G T C C T
G C G C C A G C
A
G
T

Fig. 2A

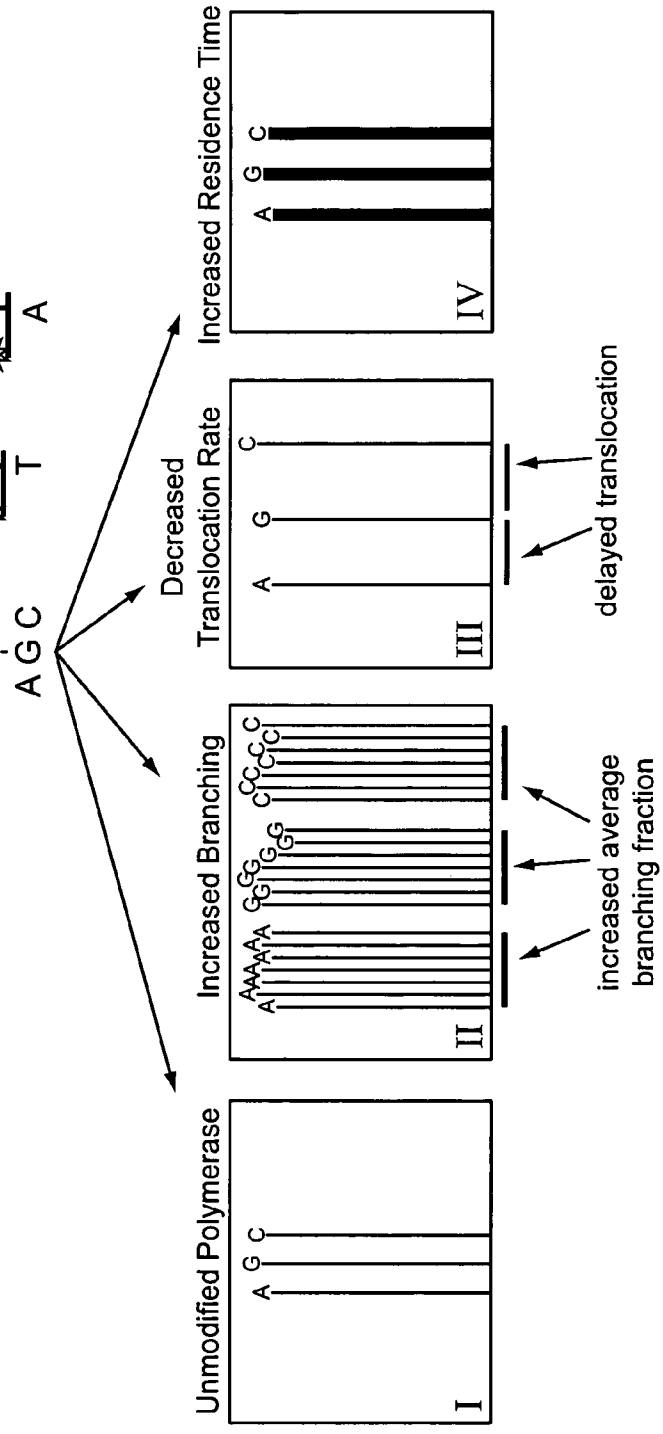

AGC
Unmodified Polymerase
I
Increased Branching
II
Decreased Translocation Rate
III
Increased Residence Time
IV
increased average branching fraction
delayed translocation

Fig. 2B

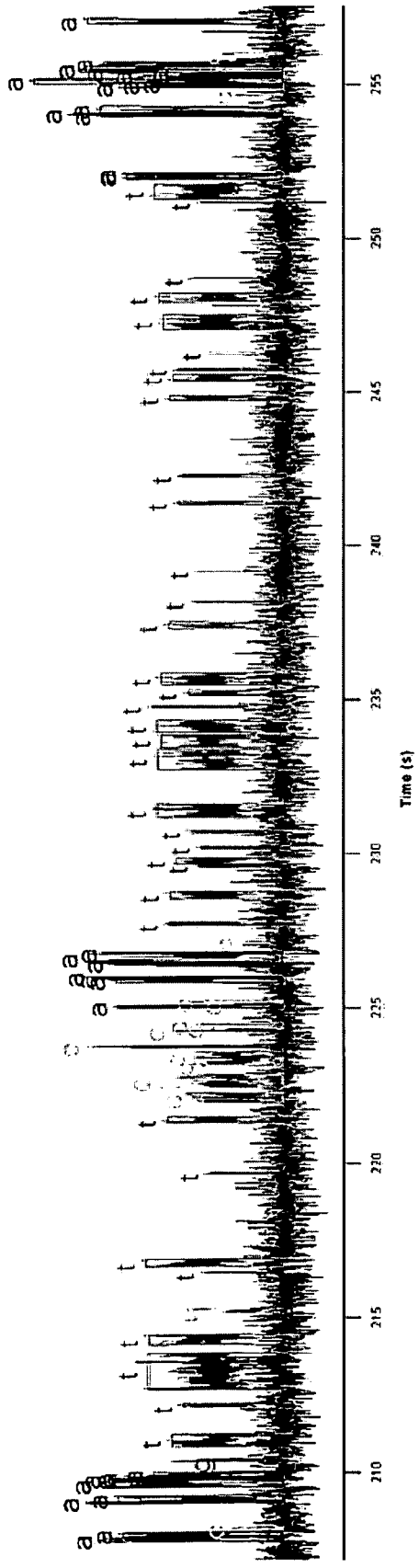


Fig. 3

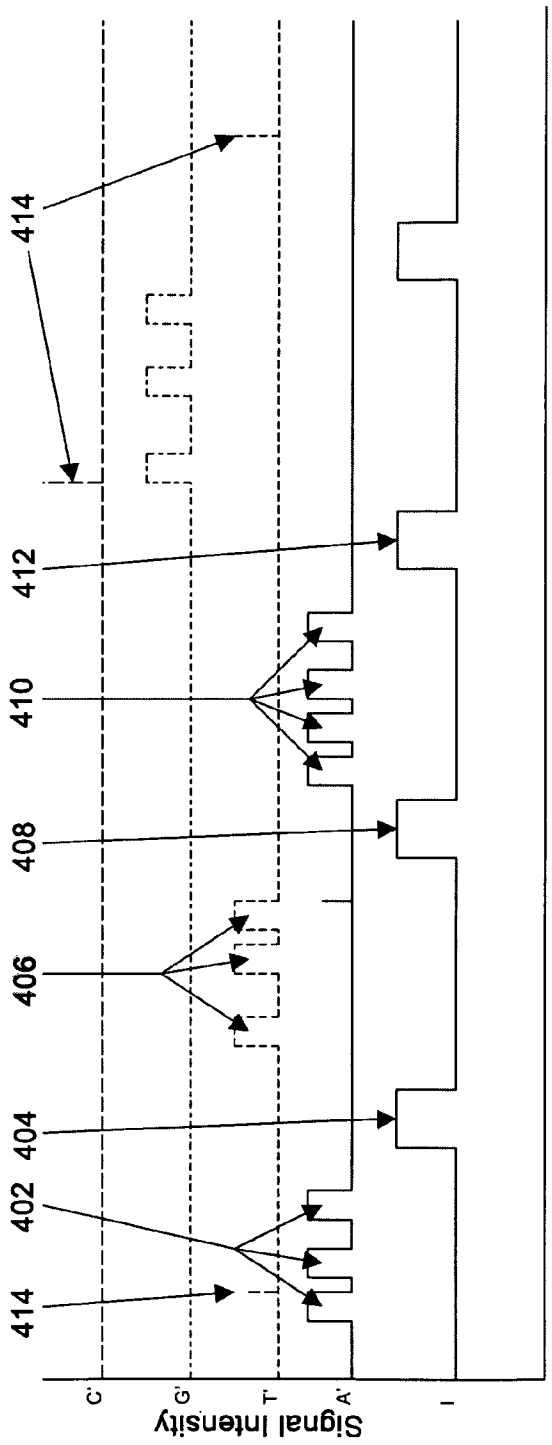


Fig. 4

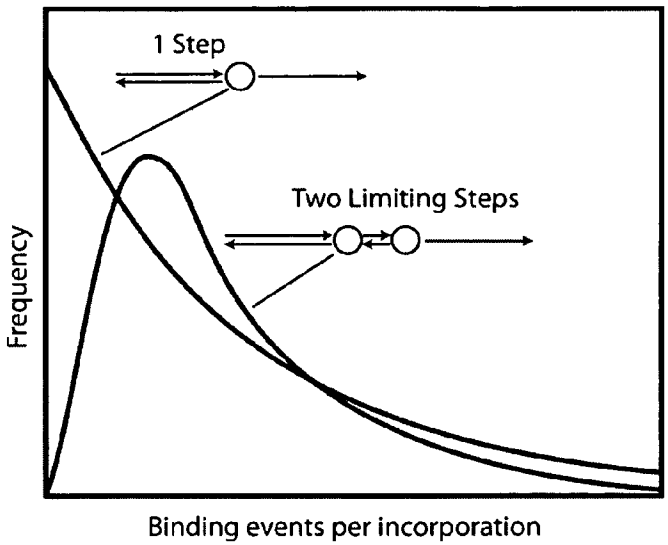

1 Step
Two Limiting Steps
Frequency
Binding events per incorporation

Fig. 5

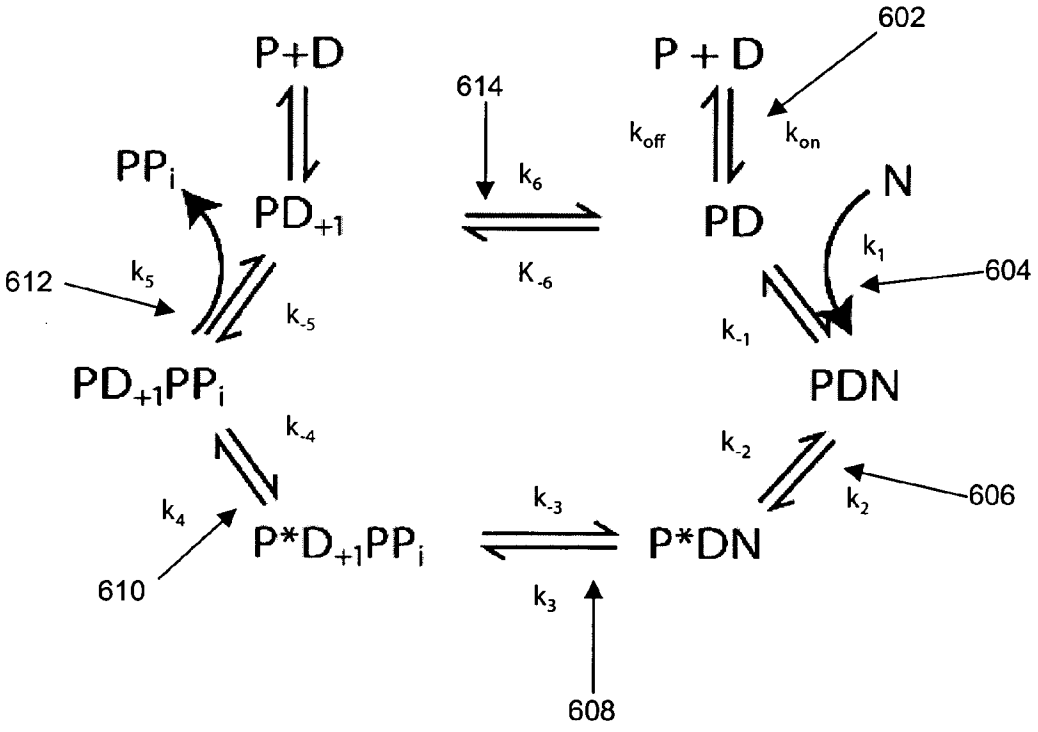

P+D
PP$_i$
PD$_{+1}$
612
k_5
k_{-5}
PD$_{+1}$PP$_i$
k_{-4}
k_4
610
P*D$_{+1}$PP$_i$
614
k_6
K_{-6}
P + D
602
k_{off}
k_{on}
N
PD
k_1
604
k_{-1}
PDN
k_{-2}
k_2
606
k_{-3}
P*DN
k_3
608

Fig. 6

7/7

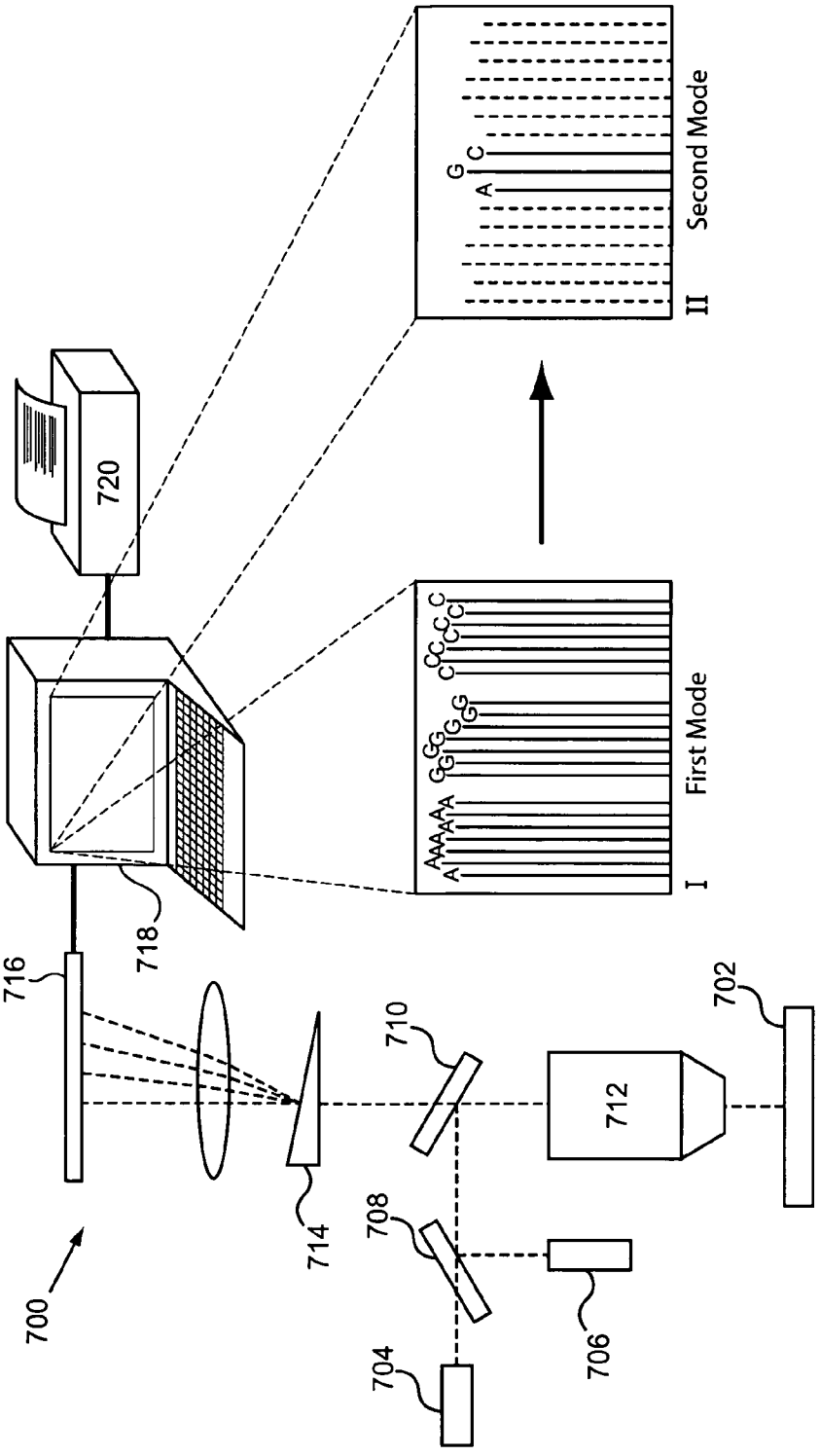

720
718
716
714
710
712
702
708
704
706
700
First Mode
I
Second Mode
II

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

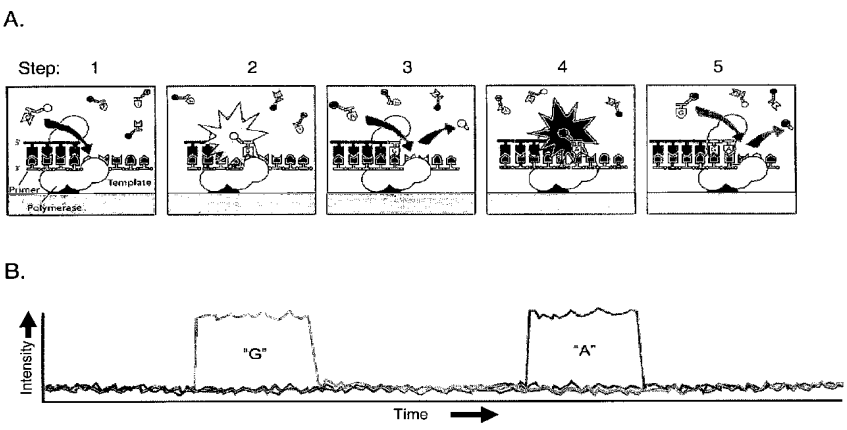


Figure 1

(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.

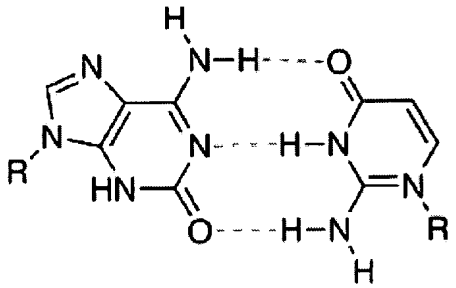
IV. Modified and Non-natural Nucleotide Analogs and Base Pairing

[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.



iso-G *iso*-C

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

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

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.

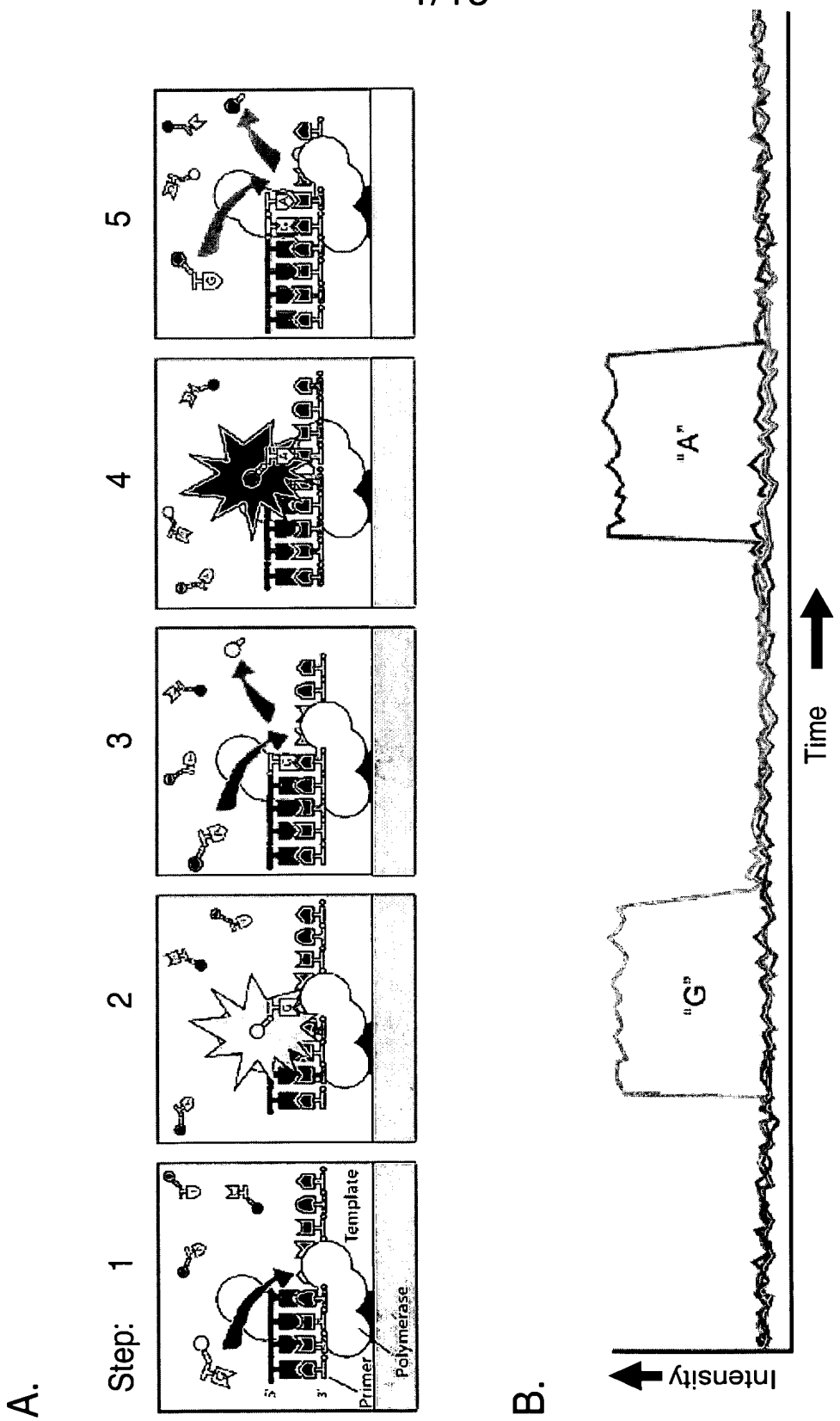

A.
Step: 1
2
3
4
5
Template
Primer
Polymerase
B.
"G"
"A"
Intensity
Time

Figure 1

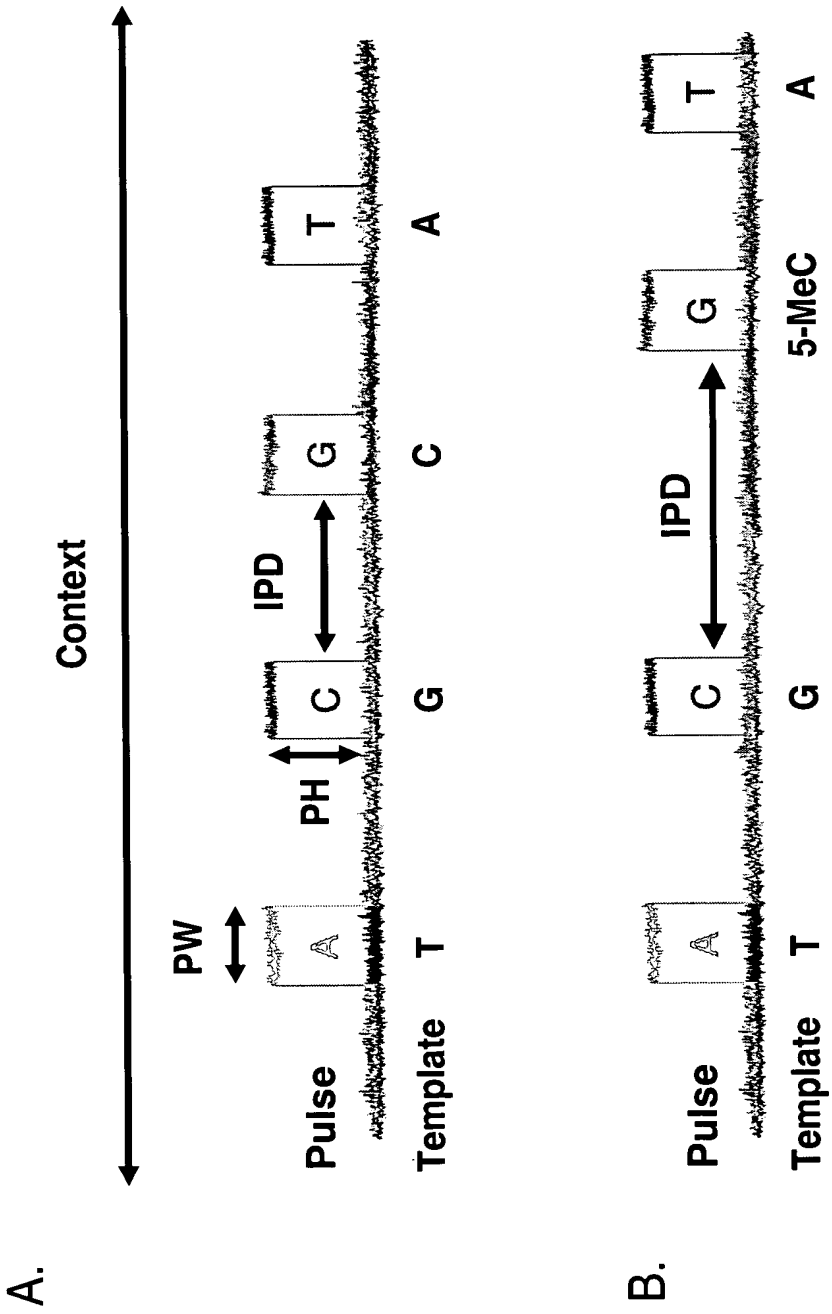

A.
Context
PW
PH
IPD
Pulse
Template
T
A
C
G
T
A
C
G
B.
IPD
Pulse
Template
T
A
C
G
5-MeC
T
A
C
G

Figure 2

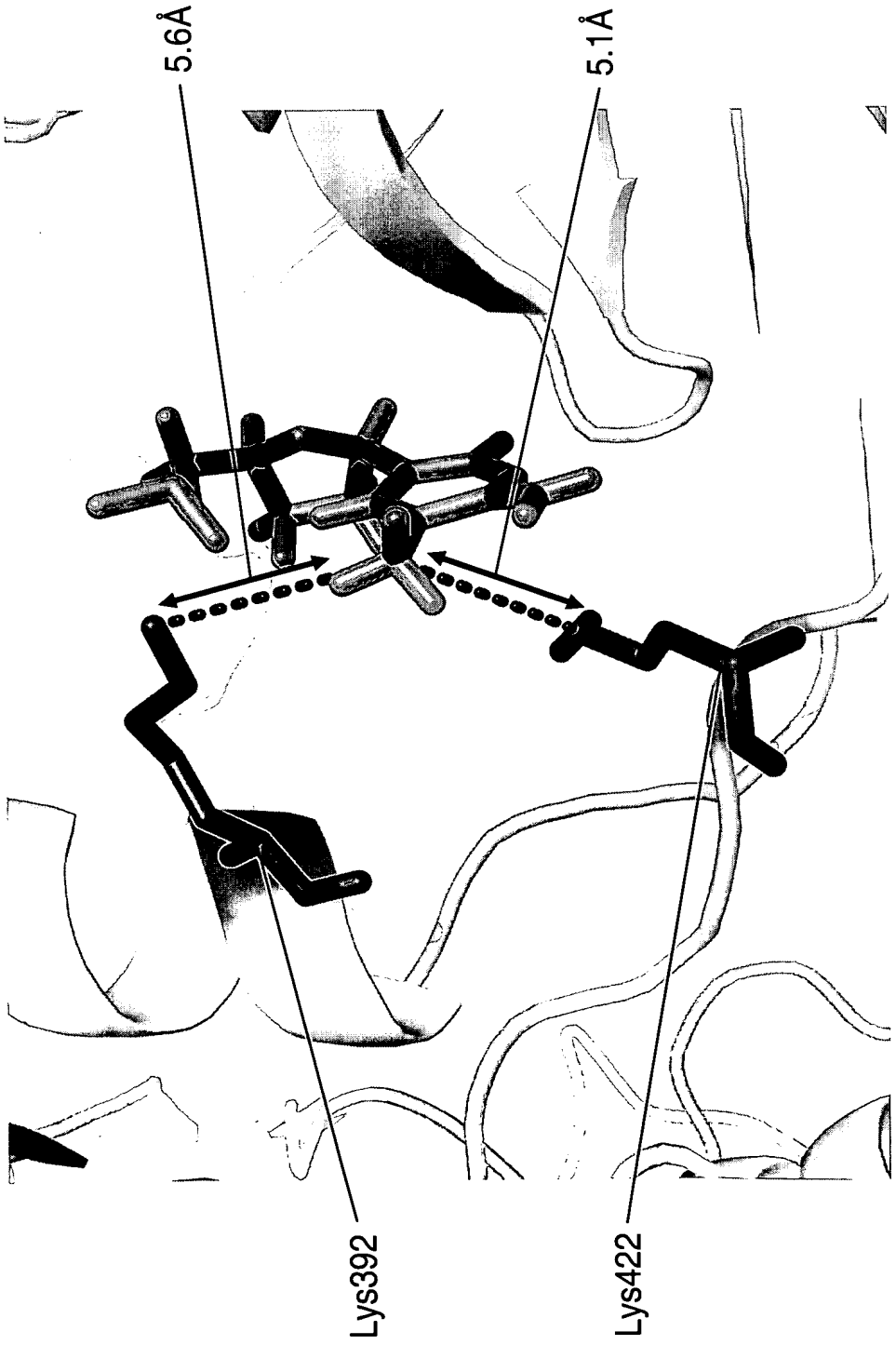


Figure 3

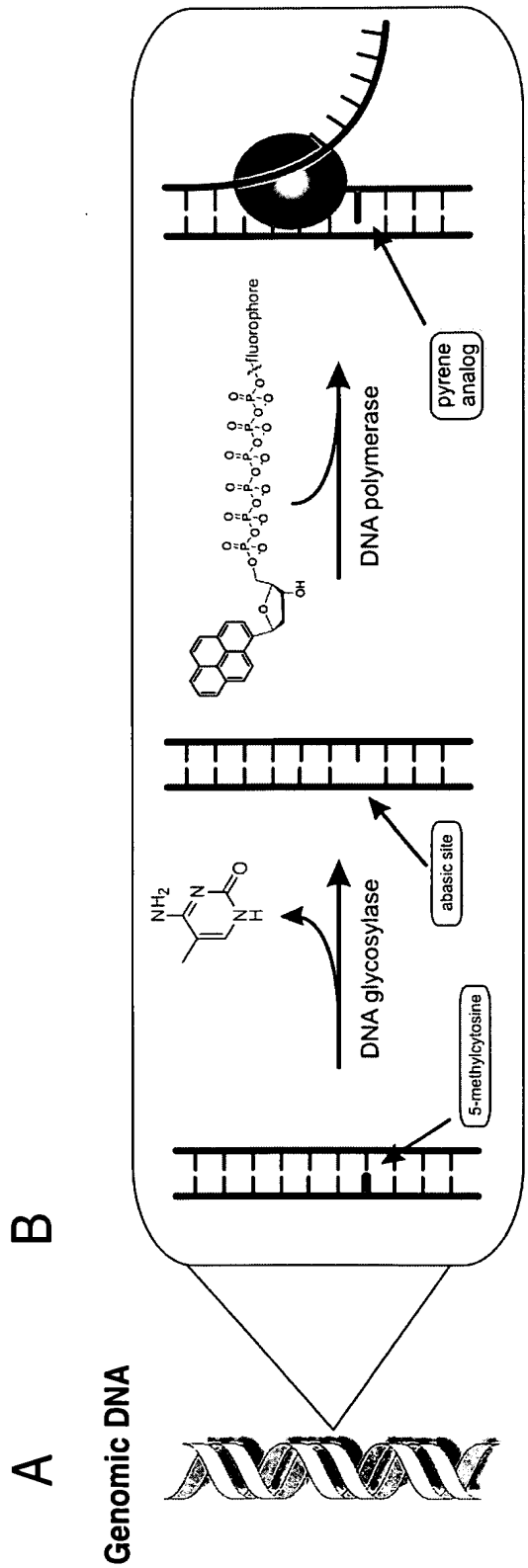

A
Genomic DNA
B
5-methylcytosine
DNA glycosylase
abasic site
DNA polymerase
pyrene analog
γ-fluorophore

Figure 4

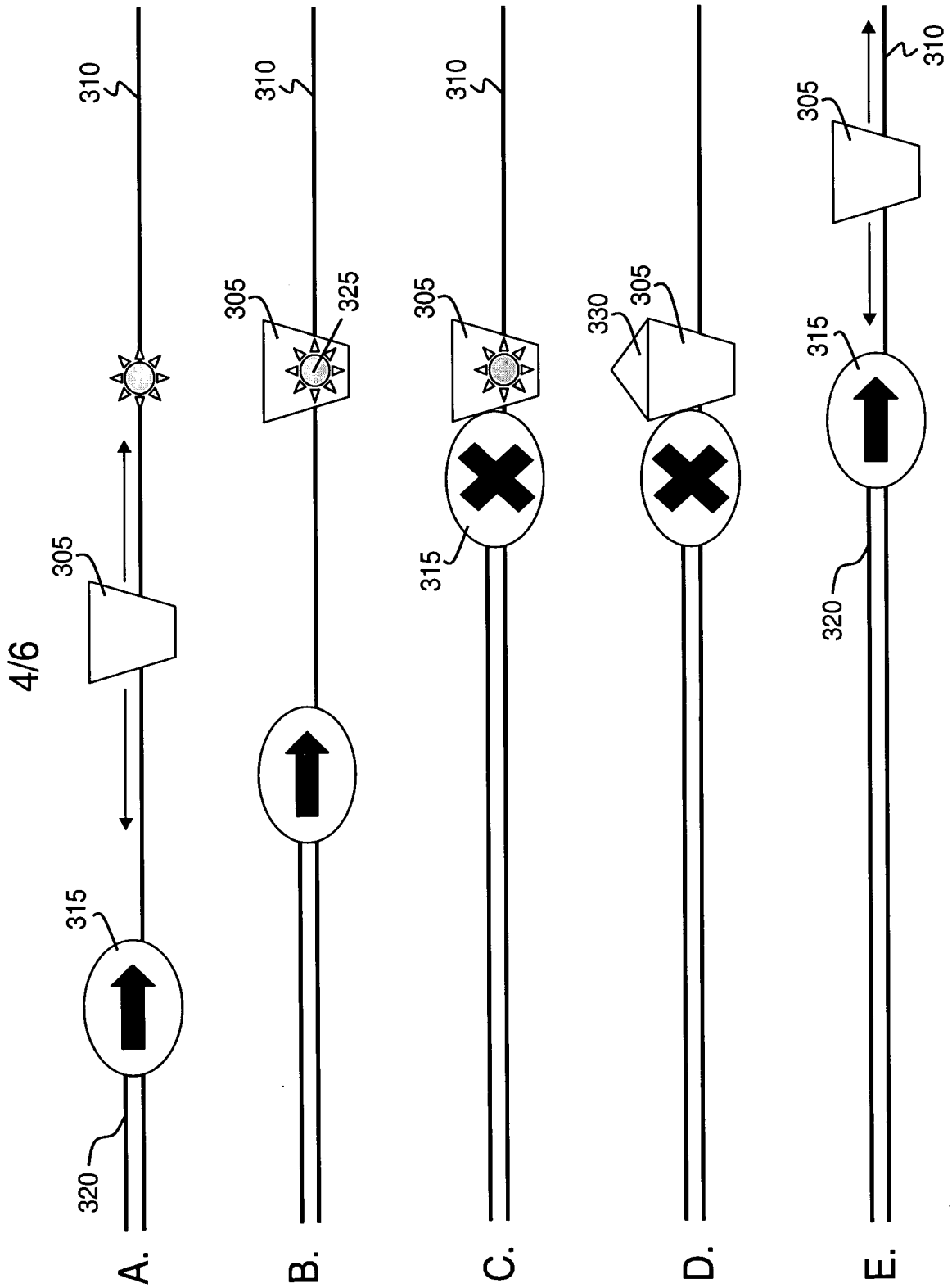


Figure 5

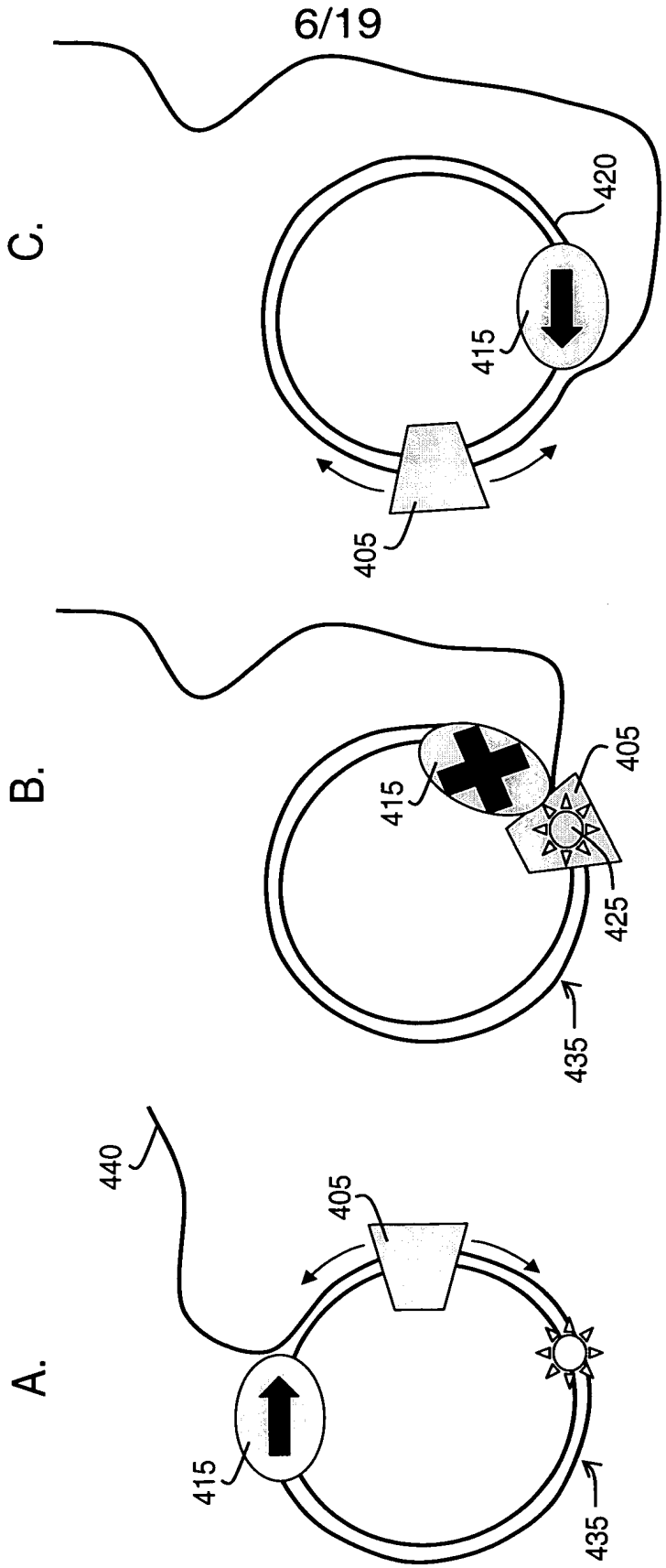


Figure 6

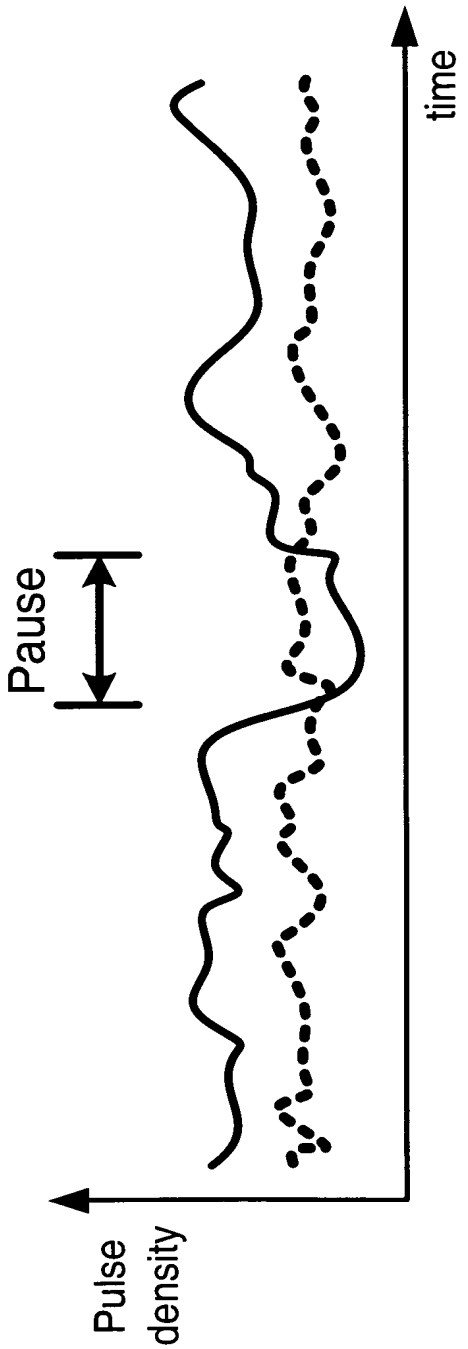


Figure 7

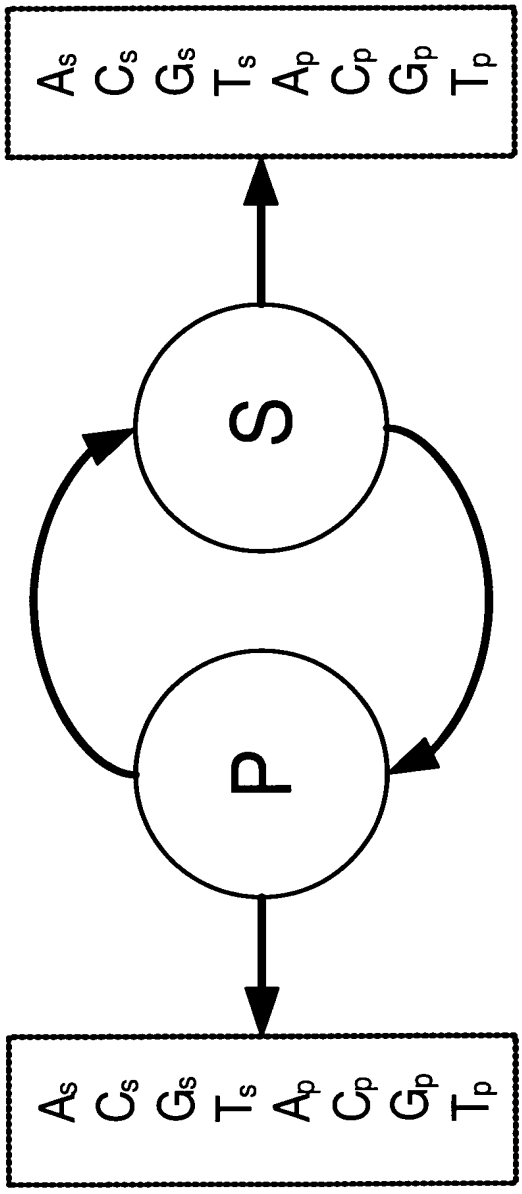


Figure 8

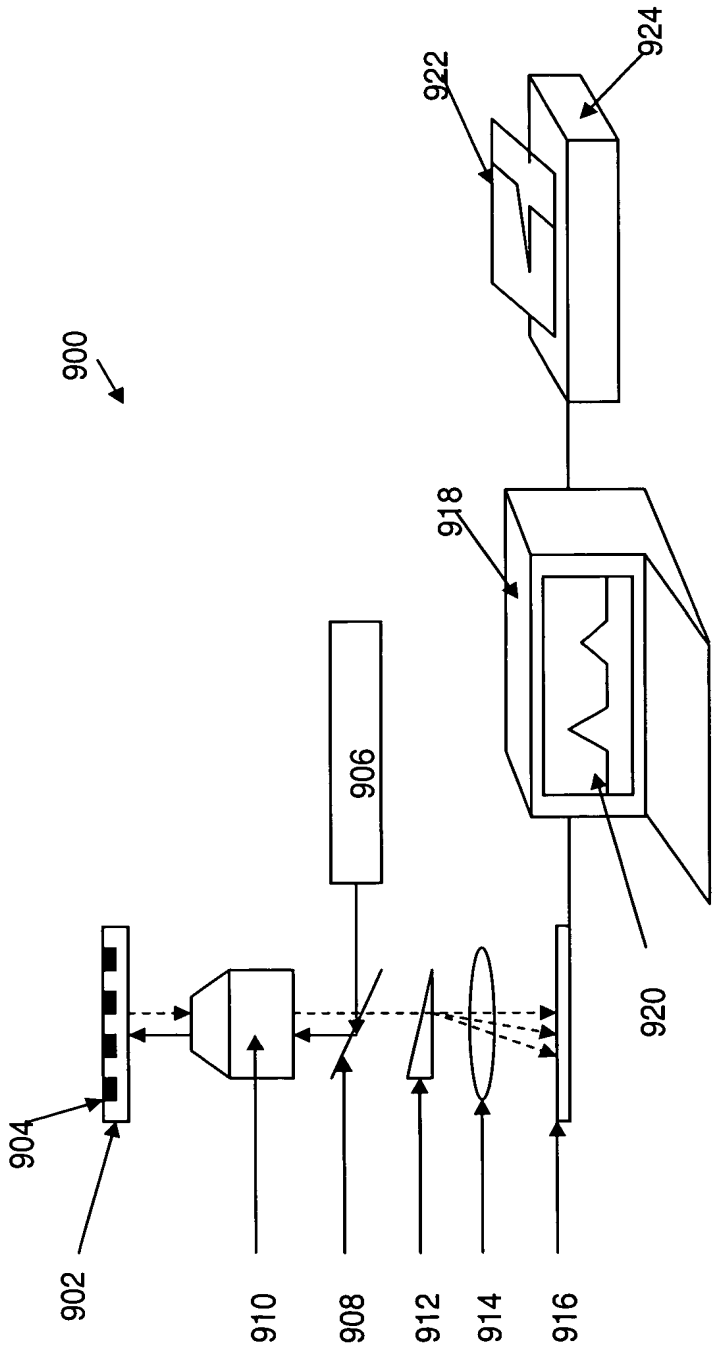


Figure 9

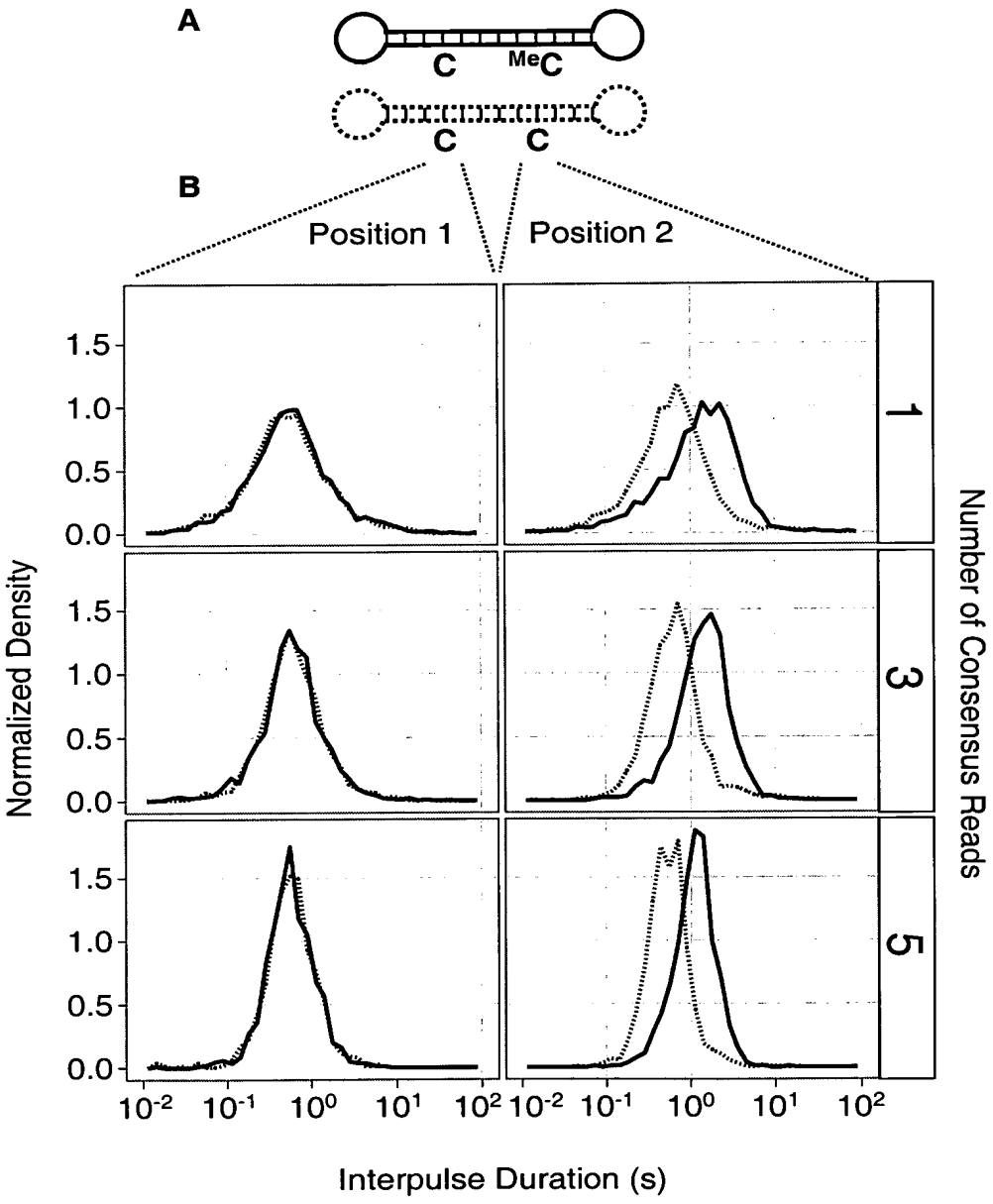

A
C
MeC
C
C
B
Position 1
Position 2
1.5
1.0
0.5
0.0
1
3
5
Normalized Density
Number of Consensus Reads
10^{-2} 10^{-1} 10^{0} 10^{1} 10^{2}
Interpulse Duration (s)

Figure 10

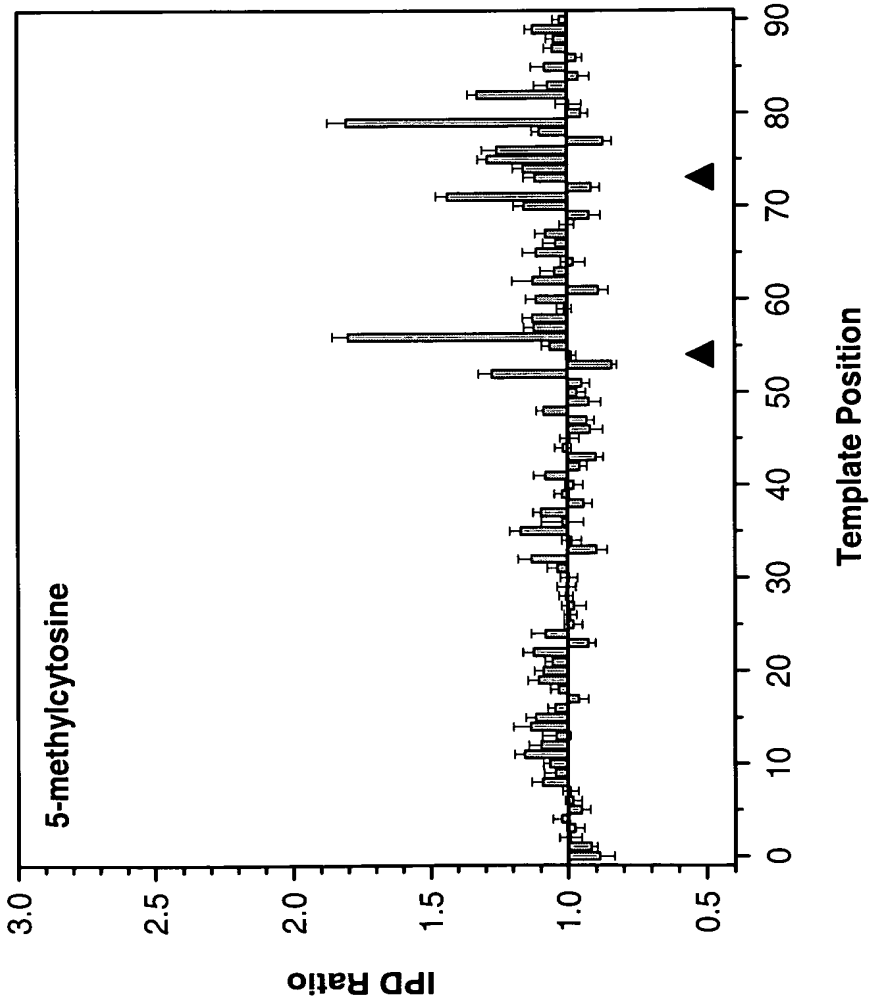

5-methylcytosine
IPD Ratio
3.0
2.5
2.0
1.5
1.0
0.5
0
10
20
30
40
50
60
70
80
90
Template Position

Figure 11

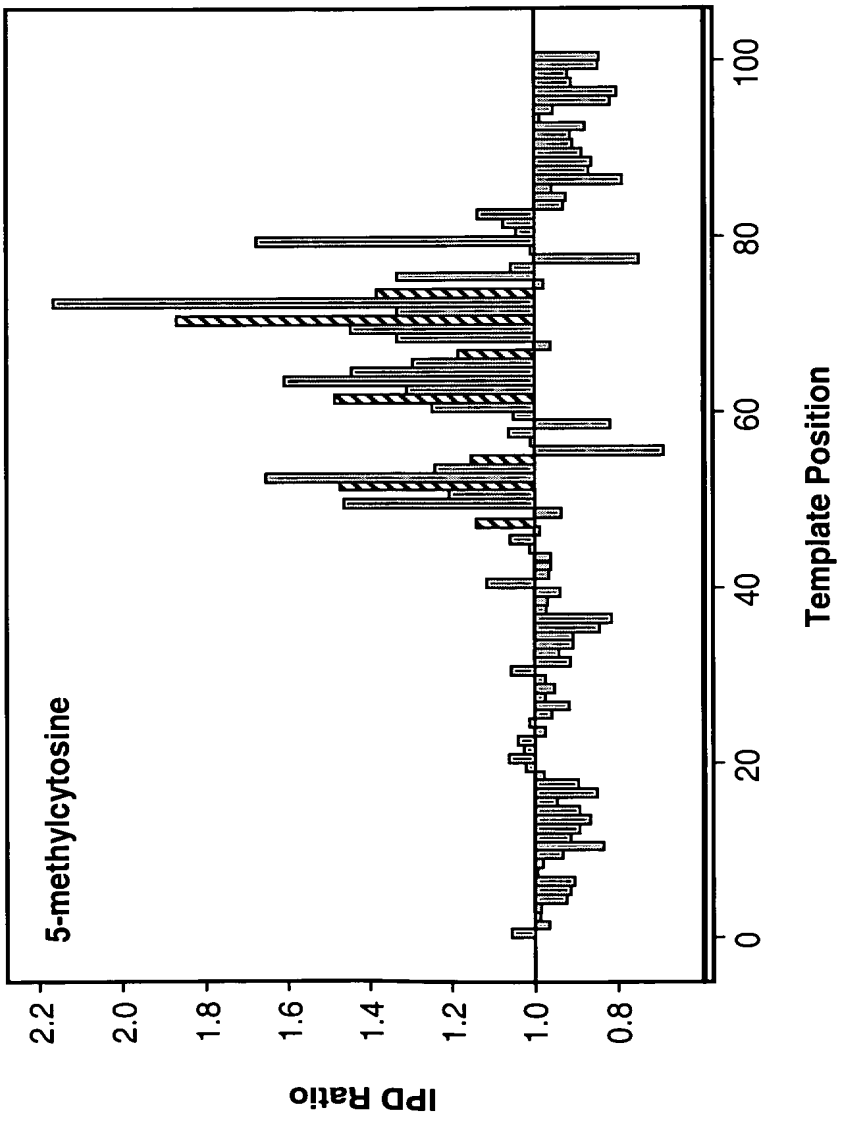

5-methylcytosine
IPD Ratio
2.2
2.0
1.8
1.6
1.4
1.2
1.0
0.8
0
20
40
60
80
100
Template Position

Figure 12

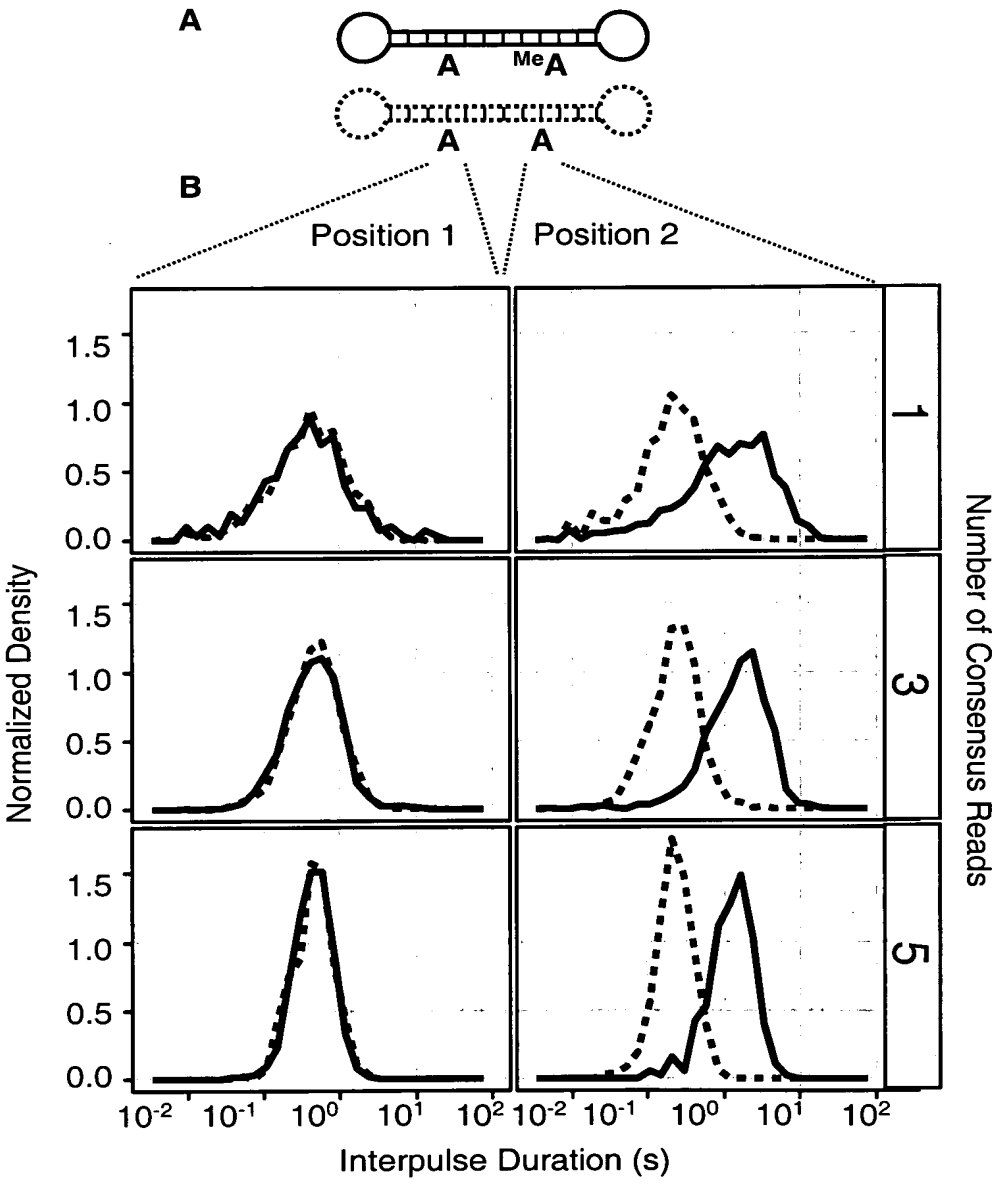

A
A
Me
A
A
A
B
Position 1
Position 2
1.5
1.0
0.5
0.0
1
3
5
Normalized Density
Number of Consensus Reads
10^{-2} 10^{-1} 10^{0} 10^{1} 10^{2}
Interpulse Duration (s)

Figure 13

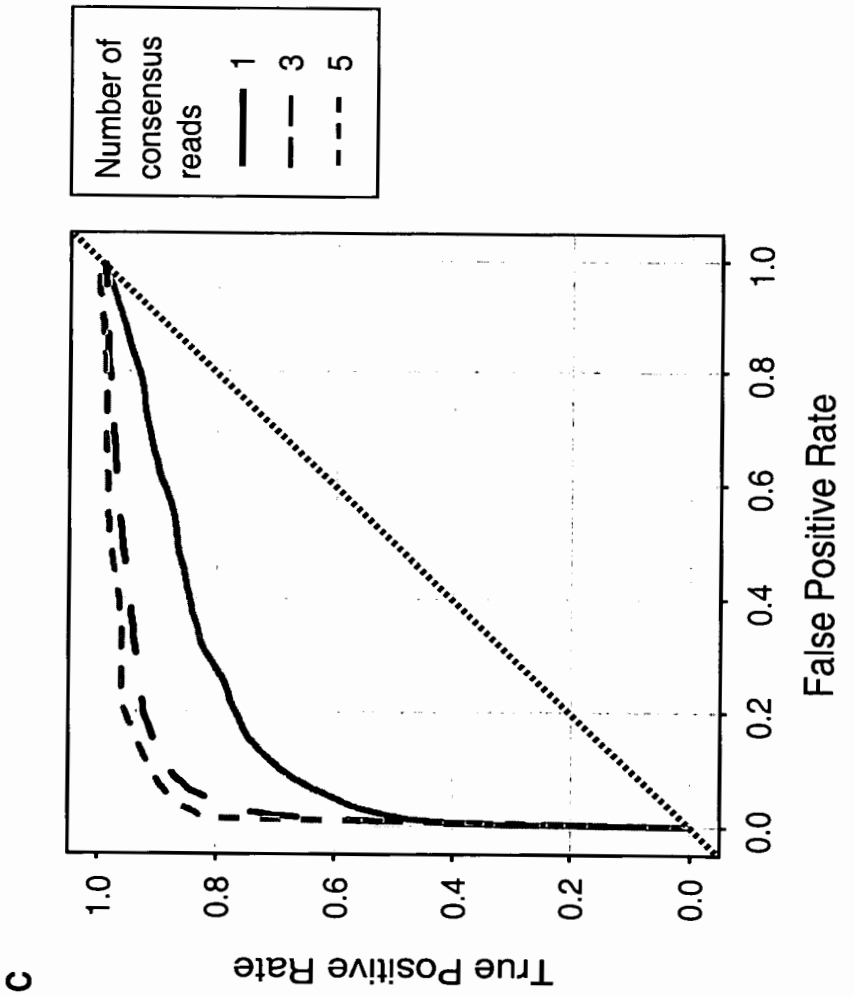


Figure 13, cont.

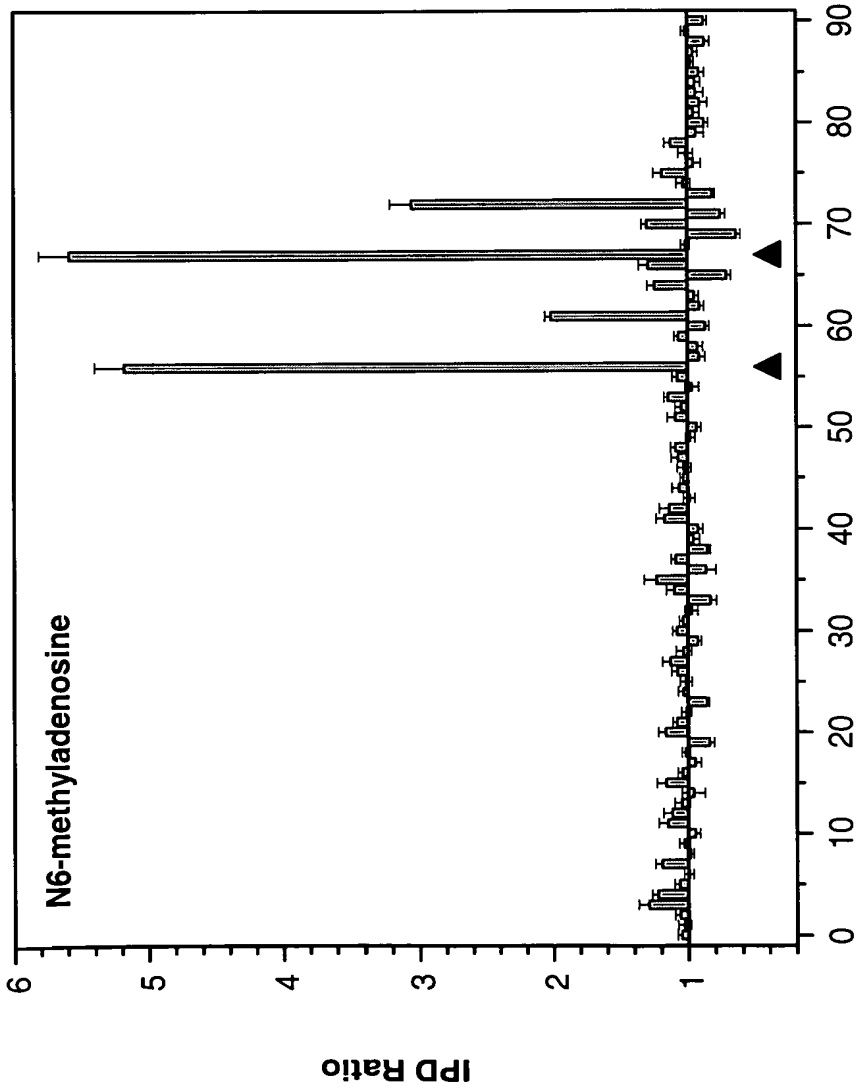

N6-methyladenosine
IPD Ratio
Template Position
0
10
20
30
40
50
60
70
80
90
1
2
3
4
5
6

Figure 14

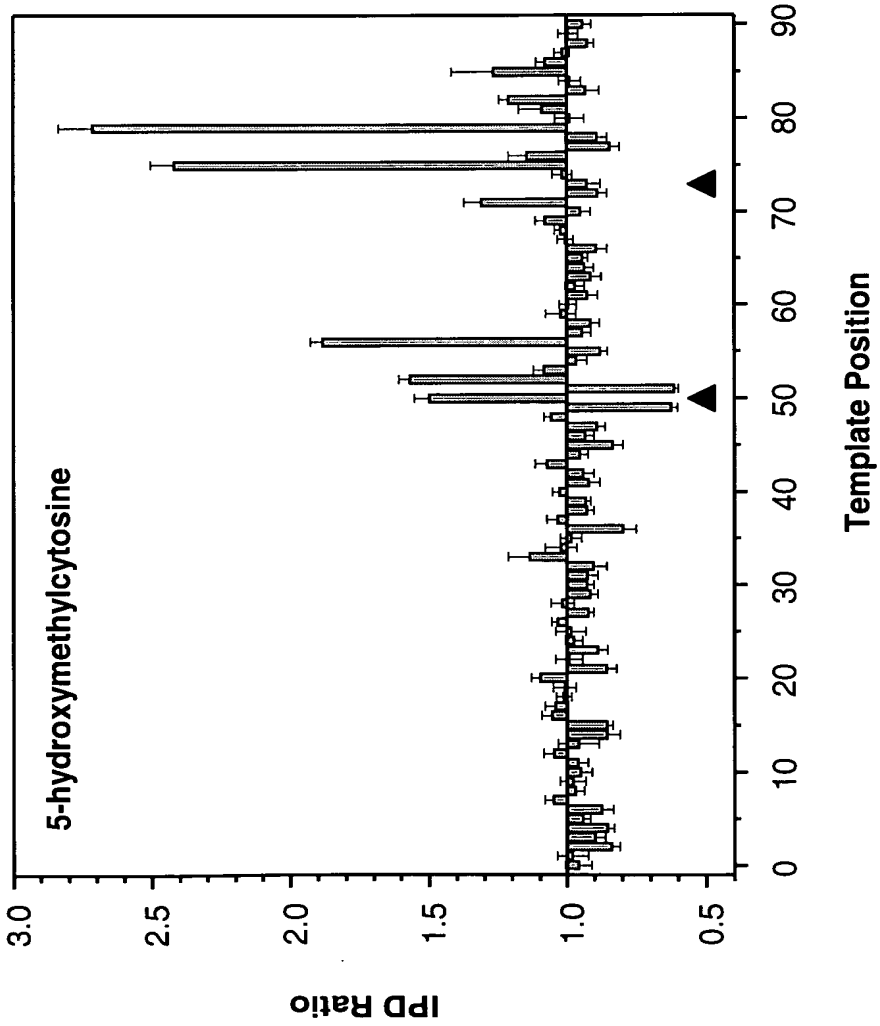

5-hydroxymethylcytosine
IPD Ratio
3.0
2.5
2.0
1.5
1.0
0.5
0
10
20
30
40
50
60
70
80
90
Template Position

Figure 15

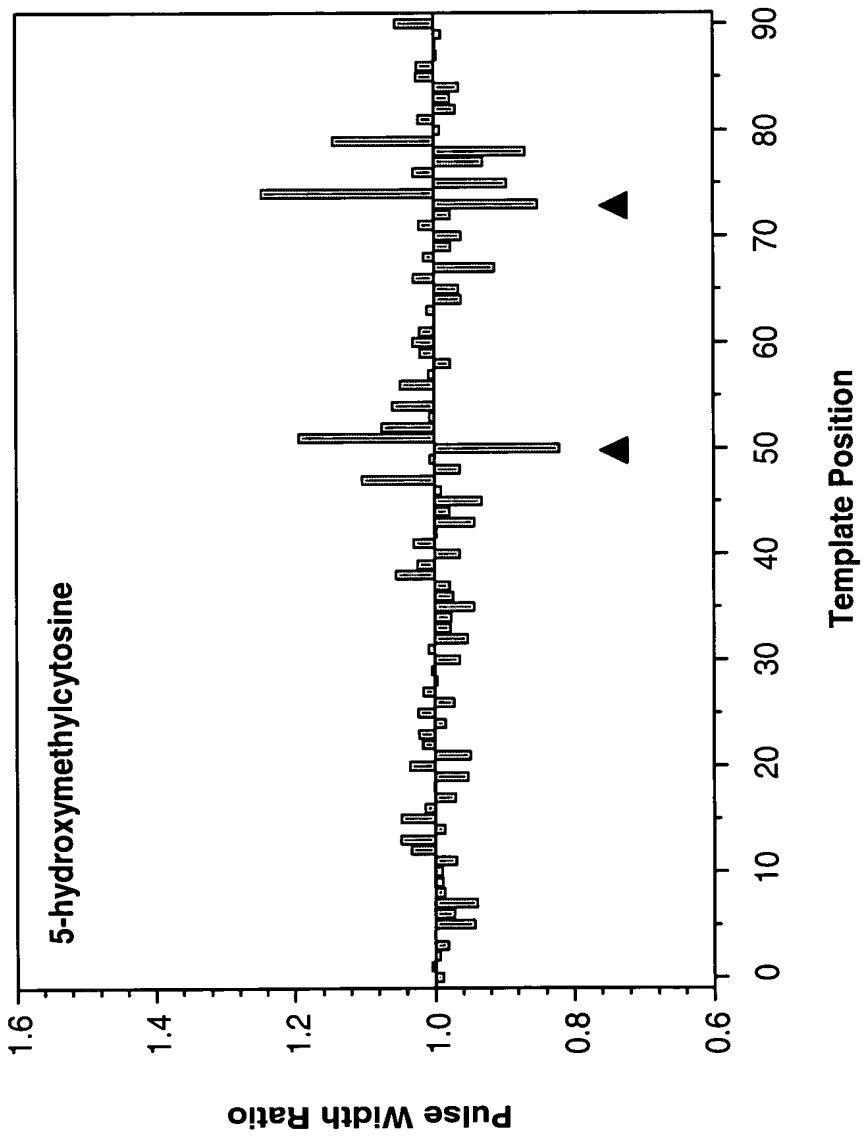

5-hydroxymethylcytosine
Pulse Width Ratio
1.6
1.4
1.2
1.0
0.8
0.6
0
10
20
30
40
50
60
70
80
90
Template Position

Figure 16

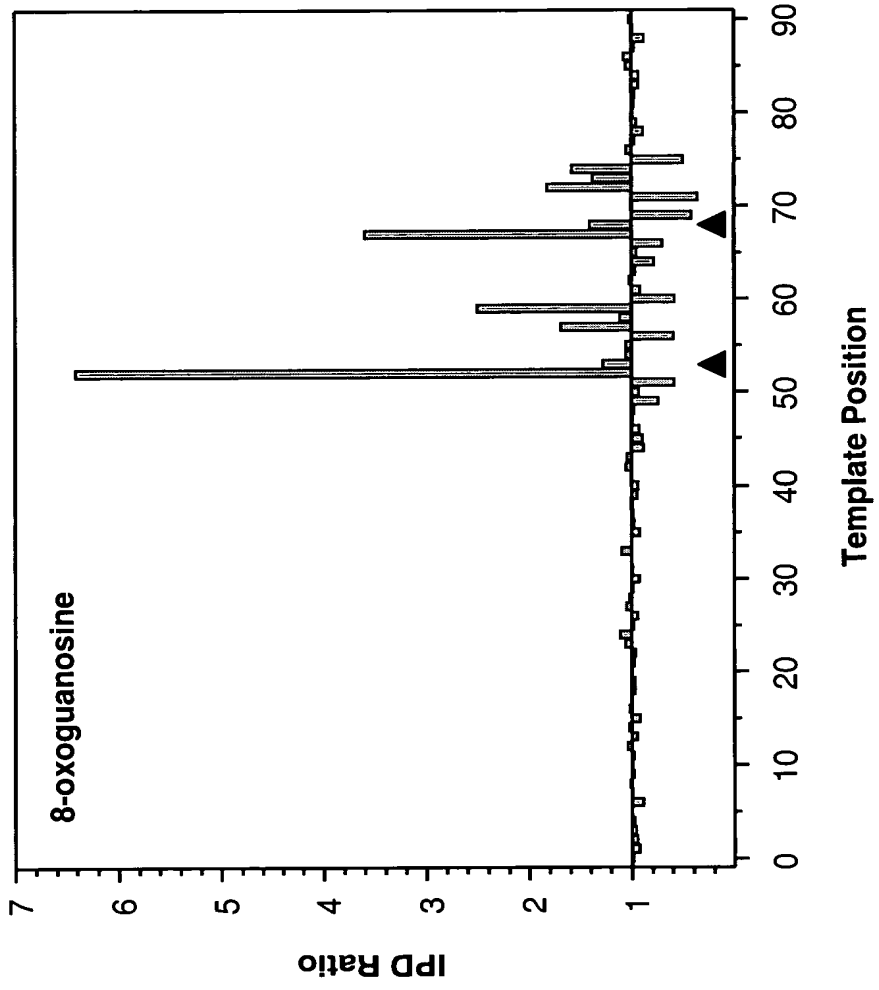

8-oxoguanosine
IPD Ratio
Template Position
1
2
3
4
5
6
7
0
10
20
30
40
50
60
70
80
90

Figure 17

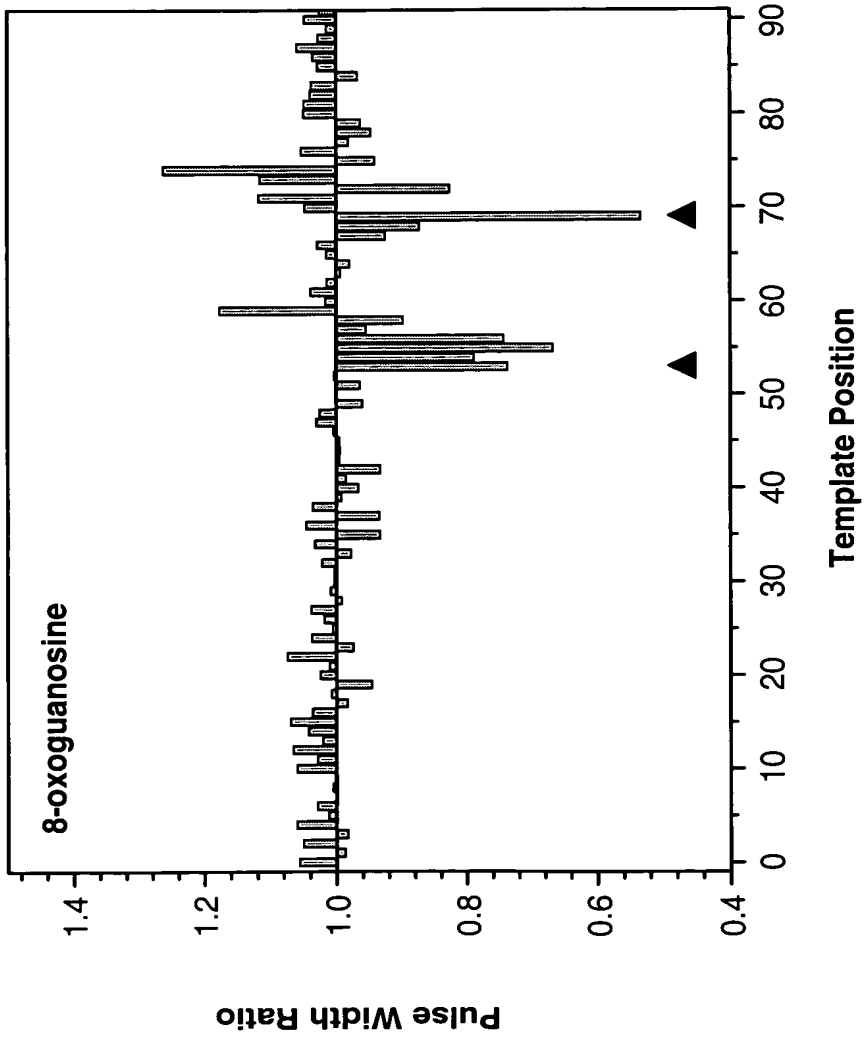

8-oxoguanosine
Pulse Width Ratio
1.4
1.2
1.0
0.8
0.6
0.4
0
10
20
30
40
50
60
70
80
90
Template Position

Figure 18

Espacenet

Bibliographic data: 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: default C12Q1/6809 (CN); C12Q1/6858 (EP, KR, US); C12Q1/6869 (EP, KR, US); C12Q1/6883 (US); G16B20/00 (EP, KR, US); G16B20/10 (EP, CN, KR, US); G16B20/20 (EP, CN, KR, US); G16B30/00 (EP, KR, US); G16B30/10 (EP, CN, KR, US); G16B30/20 (EP, CN, KR, US); G16B40/00 (KR); G16H50/20 (EP, KR, US); G16H50/30 (EP, KR, US); C12Q2600/118 (US); C12Q2600/156 (US)

C-sets

- C12Q1/6809, C12Q2535/122, C12Q2537/165 **(CN)**;
- C12Q1/6858, C12Q2535/122, C12Q2537/165 **(EP, US)**;
- C12Q1/6869, C12Q2535/122, C12Q2537/165 **(EP, US)**

less

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 蒋会娟 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)

(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

(12) STANDARD PATENT (11) Application No. **AU 2012380221 B2**
(19) AUSTRALIAN PATENT OFFICE

(54) Title
Method, system and computer readable medium for determining base information in predetermined area of fetus genome

(51) International Patent Classification(s)
C12Q 1/68 (2006.01)

(21) Application No: **2012380221** (22) Date of Filing: **2012.05.14**

(87) WIPO No: **WO13/170429**

(43) Publication Date: **2013.11.21**
(44) Accepted Journal Date: **2016.09.29**

(71) Applicant(s)
BGI Diagnosis Co., Ltd.

(72) Inventor(s)
Chen, Shengpei;Ge, Huijuan;Li, Xuchao;Yi, Shang;Wang, Jian;Wang, Jun;Yang, Huanming;Zhang, Xiuqing

(74) Agent / Attorney
Shelston IP Pty Ltd., Level 21, 60 Margaret Street, Sydney, NSW, 2000

(56) Related Art
WO 2007/092473
US 2012/095697

(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

2012380221 01 Sep 2016

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

TECHNICAL FIELD

Embodiments of the present disclosure generally relate to a method of determining base information of a predetermined region in a fetal genome, and a system and a computer readable medium thereof.

BACKGROUND

Any discussion of the prior art throughout the specification should in no way be considered as an admission that such prior art is widely known or forms part of common general knowledge in the field.

Genetic diseases are one kind of diseases caused by changes of genetic materials, having characteristics of being congenital, familial, permanent and hereditary. The genetic diseases may be categorized into 3 classes: monogenetic disease, polygenetic disorder and chromosome abnormality. In which the monogenetic disease is mostly because of genetic function abnormality caused by dominant or recessive inheritance of a single disease-causing gene; while the polygenetic disorder is a kind of disease caused by a plurality of gene changes, which may be influenced by external environment to some extent; and the chromosome abnormality includes number abnormality and structure abnormality, with a most common example being as a Down's Syndrome resulting from Trisomy 21, of which a child patient presenting congenital traits such as mongolism and abnormal body shape, etc. Since there are no effective therapeutic treatments for genetic diseases so far, it can only pertinently perform supportive treatments or drug remission with expensive cost, which may bring heavy burdens both in economy and spirit for societies and families. Thus, it is extremely necessary to do some preventive work by detecting pathological status with a fetus before birth, to achieve a purpose of good prenatal and postnatal care.

However, related detection method still needs to be improved.

2012380221 01 Sep 2016

Unless the context clearly requires otherwise, throughout the description and the claims, the words "comprise", "comprising", and the like are to be construed in an inclusive sense as opposed to an exclusive or exhaustive sense; that is to say, in the sense of "including, but not limited to".

SUMMARY

According to a first aspect, the invention provides a method of determining base information of a predetermined region in a fetal genome, comprising following steps:

constructing a sequencing library based on a genomic DNA sample of a fetus;

subjecting the sequencing library to sequencing, to obtain a sequencing result of the fetus consisting of a plurality of sequencing data; and

determining the base information of the predetermined region based on the sequencing result of the fetus combining with genetic information of a related individual using a hidden Markov Model,

wherein the base information of the predetermined region comprises the most possible fetal haplotype recombination,

and the base information of the predetermined region is obtained by:

constructing a probability distribution vector of an initial state and a transition matrix of haplotypes recombination;

constructing a probability matrix of observations;

constructing a partial probability matrix, and a reversal cursor;

determining a final state, and tracing back an optional path; and

outputting a result to analyze the DNA sequence of the fetal genome.

According to a second aspect, the invention provides a system for determining base information of a predetermined region in a fetal genome, comprising:

a library constructing apparatus, adapted for constructing sequencing library based on a genomic DNA sample of a fetus;

a sequencing apparatus, connected to the library constructing apparatus, and adapted for subjecting the sequencing library to sequencing, to obtain a sequencing result of the fetus consisting of a plurality of sequencing data; and

an analyzing apparatus, connected to the sequencing apparatus, and adapted for determining the base information of the predetermined region based on the

sequencing result of the fetus combining with genetic information of a related individual using a hidden Markov Model,

wherein the base information of the predetermined region comprises the most possible fetal haplotype recombination,

and the base information of the predetermined region is obtained by:

constructing a probability distribution vector of an initial state and a transition matrix of haplotypes recombination;

constructing a probability matrix of observations;

constructing a partial probability matrix, and a reversal cursor;

determining a final state, and tracing back an optional path; and

outputting a result to analyze the DNA sequence of the fetal genome.

According to a third aspect, the invention provides a computer readable medium comprising a plurality of instructions, adapted for determining base information of a predetermined region based on a sequencing result of a fetus combining with genetic information of a related individual using a hidden Markov Model, which causes the apparatus to perform:

constructing a sequencing library based on the genomic DNA sample of the fetus;

subjecting the sequencing library to sequencing, to obtain a sequencing result of the fetus consisting of a plurality of sequencing data; and

determining the base information of the predetermined region based on the sequencing result of the fetus combining with genetic information of a related individual using the hidden Markov Model,

wherein the base information of the predetermined region comprises the most possible fetal haplotype recombination,

and the base information of the predetermined region is obtained by:

constructing a probability distribution vector of an initial state and a transition matrix of haplotypes recombination;

constructing a probability matrix of observations;

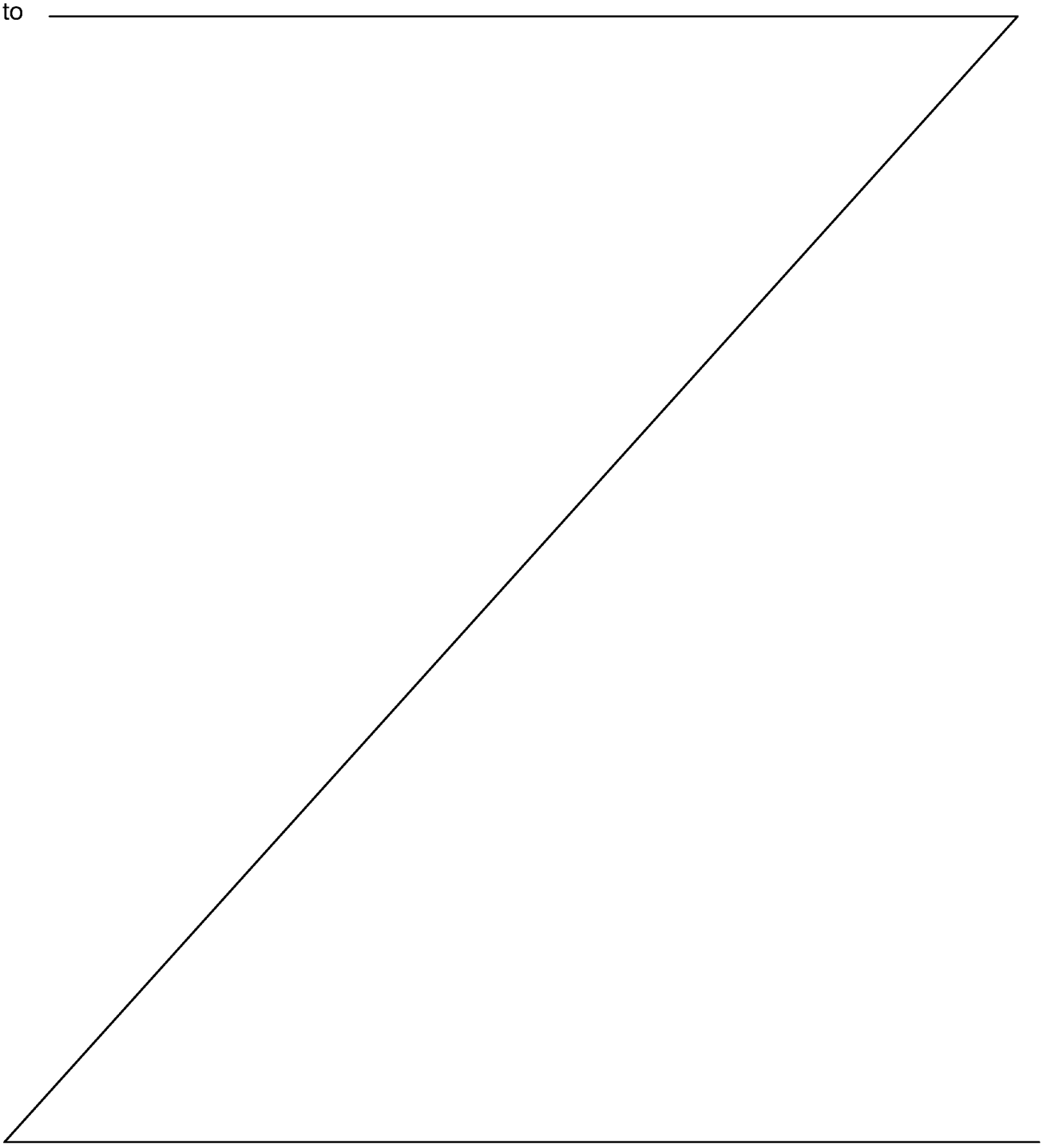
constructing a partial probability matrix, and a reversal cursor;

determining a final state, and tracing back an optional path; and

outputting a result to analyze the DNA sequence of the fetal genome.

Embodiments of the present disclosure seek to solve at least one of the problems existing in the related art to at least some extent, or to provide a useful alternative.

Embodiments of a first broad aspect of the present disclosure provide a method of determining base information of a predetermined region in a fetal genome. According to



1c