[0061] As used herein, "administering" or "administration of" a compound or an agent to a subject can be carried out using one of a variety of methods known to those skilled in the art. For example, a compound or an agent can be administered, intravenously, arterially, intradermally, intramuscularly, intraperitonealy, intravenously, subcutaneously, ocularly, sublingually, orally (by ingestion), intranasally (by inhalation), intraspinally, intracerebrally, and transdermally (by absorbtion, e.g., through a skin duct). A compound or agent can also appropriately be introduced by rechargeable or biodegradable polymeric devices or other devices, e.g., patches and pumps, or formulations, which provide for the extended, slow, or controlled release of the compound or agent. Administering can also be performed, for example, once, a plurality of times, and/or over one or more extended periods. In some aspects, the administration includes both direct administration, including self-administration, and indirect administration, including the act of prescribing a drug. For example, as used herein, a physician who instructs a patient to self-administer a drug, or to have the drug administered by another and/or who provides a patient with a prescription for a drug is administering the drug to the patient. In some embodiments, a compound or an agent is administered orally, e.g., to a subject by ingestion, or intravenously, e.g., to a subject by injection. In some embodiments, the orally administered compound or agent is in an extended release or slow release formulation, or administered using a device for such slow or extended release.

[0062] In certain embodiments, markers used in the methods of invention are up-regulated or activated in the combination sample compared to the phagocytic cells, =2n phagocytic cells, non-phagocytic cells, or control cells that are substantially free of cells affected by the disease or condition. Different diseases or conditions can be associated with either up-regulation (or activation) or down-regulation (or inhibition) of different markers. As used herein, "up-regulation or up-regulated" can refer to an increase in expression levels (e.g., gene expression or protein expression), gene copy numbers, gene dosages, and other qualitative or quantitative detectable state of the markers. Similarly, "down-regulation or down-regulated" can refer to an increase in expression levels, gene copy numbers, gene dosages, and other qualitative or quantitative detectable state of the markers. In certain embodiments, markers used in the methods of invention are down-regulated or inhibited in the combination sample

compared to the phagocytic cells, =2n phagocytic cells, non-phagocytic cells, or control cells that are substantially free of cells affected by the disease or condition. Different diseases or conditions can be associated with either up-regulation (or activation) or down-regulation (or inhibition) of different markers. As used herein, "activation or activated" can refer to an active state of the marker, e.g., a phosphorylation state, a DNA methylation state, or a DNA acetylation state. Similarly, "inhibition or inhibited" can refer to a repressed state or an inactivated state of the marker, e.g., a de-phosphorylation state, a ubiquitination state, a DNA de-methylation state.

[0063] In certain embodiments, methods of this invention can comprise extracting or enriching markers from cell-free bodily fluids. Any known extraction and enrichment methods can be used herein. In certain embodiments, methods of this invention also comprise at least one of the following steps before determination of various profiles: i) lysing the =2n phagocytic, non-phagocytic, or control cells; and ii) extracting cellular contents from the lysed cells. Any known cell lysis and extraction methods can be used herein. In certain embodiments, at least one or more markers of a disease or condition are present in the combination sample. In certain embodiments, there is no marker present in the cellular contents of the non-phagocytic, =2n phagocytic, or control cells.

[0064] In certain embodiments, methods of this invention also comprise at least one of the following steps before determination of various profiles: i) lysing the phagocytic, >2n phagocytic cells, or =2n phagocytic cells; and ii) extracting cellular contents from the lysed phagocytic, >2n phagocytic, or =2n phagocytic cells. In certain embodiments, the cellular contents of the phagocytic or >2n phagocytic cells comprise various types of materials that they have engulfed, such as viable diseased cells, dead diseased cells, apoptotic diseased cells, circulating tumor cells, infectious agents, fetal cells, trophoblasts, or fragments thereof. In certain embodiments, at least one or more markers of a disease or condition are present in the cellular contents of the phagocytic or >2n phagocytic cells.

[0065] In certain embodiments, methods of this invention also comprise at least one of the following steps before determination of various profiles: i) lysing the circulating vesicles; and ii) extracting the contents from the lysed circulating vesicles. In certain embodiments, the contents of the circulating vesicles comprise various types of materials, such as proteins and nucleic acids. In certain embodiments, at least one or more markers of a disease or condition are present in the contents of the circulating vesicles.

[0066] In certain embodiments, methods of this invention also comprise at least one of the following steps before determination of various profiles: i) lysing the circulating diseased cells; and ii) extracting the contents from the lysed circulating diseased cells. In certain embodiments, at least one or more markers of a disease or condition are present in the contents of the circulating diseased cells.

[0067] In certain embodiments, methods of this invention further comprise comparing the identified difference of the disease or condition-specific markers to a repository of at least one markers known in the art. Such comparison can further confirm the presence of the disease or condition. In some embodiments, the repository of the known markers can be obtained by data mining. The term "data mining", as used herein, refers to a process of finding new data patterns, relations, or correlations derived from the known data of the databases and of extracting practicable information in the future. Typically a computer-based system can be trained on data to perform the data mining, e.g., to classify the input data and then subsequently used with new input data to make decisions based on the training data. These systems include, but are not limited, expert systems, fuzzy logic, non-linear regression analysis, multivariate analysis, decision tree classifiers, and Bayesian belief networks.

[0068] In certain embodiments, the phagocytic cells, >2n phagocytic cells, circulating vesicles, circulating diseased cells, control cells, and =2n phagocytic cells are isolated from a bodily fluid sample, tissues, or cells. Exemplary bodily fluid samples can be whole blood, urine, stool, saliva, lymph fluid, cerebrospinal fluid, synovial fluid, cystic fluid, ascites, pleural effusion, fluid obtained from a pregnant

woman in the first trimester, fluid obtained from a pregnant woman in the second trimester, fluid obtained from a pregnant woman in the third trimester, maternal blood, amniotic fluid, chorionic villus sample, fluid from a preimplantation embryo, maternal urine, maternal saliva, placental sample, fetal blood, lavage and cervical vaginal fluid, interstitial fluid, buccal swab sample, sputum, bronchial lavage, Pap smear sample, or ocular fluid. In some embodiments, the phagocytic cells, >2n phagocytic cells, and =2n phagocytic cells are isolated from white blood cells. In certain embodiments, the >2n phagocytic cells and the =2n phagocytic cells are separated from a population of phagocytic cells.

[0069] In certain embodiments, the cell-free bodily fluid comes from a bodily fluid sample. Exemplary bodily fluid samples can be whole blood, urine, stool, saliva, lymph fluid, cerebrospinal fluid, synovial fluid, cystic fluid, ascites, pleural effusion, fluid obtained from a pregnant woman in the first trimester, fluid obtained from a pregnant woman in the second trimester, fluid obtained from a pregnant woman in the third trimester, maternal blood, amniotic fluid, chorionic villus sample, fluid from a preimplantation embryo, maternal urine, maternal saliva, placental sample, fetal blood, lavage and cervical vaginal fluid, interstitial fluid, buccal swab sample, sputum, bronchial lavage, Pap smear sample, or ocular fluid. In some embodiments, the cell-free bodily fluids are obtained by separating cells from the bodily fluid sample by methods known in the art, such as extraction, centrifugation, and filtration.

[0070] In some embodiments, a component for use in a combination sample may be obtained by removing cells from a bodily fluid. In some embodiments, a component for use in a combination sample may be obtained by destroying (e.g., lysing) cells in a bodily fluid. These embodiments may be taken in combination, for example, by removing some populations of cells and destroying other populations of cells. In some embodiments, a sample of whole blood may be used to create a combination sample, for example, by removing red blood cells, serum, and T cells, and using the remainder as the combination sample.

[0071] In certain embodiments, tissue or fluid samples including cells having a DNA content of 2n are obtained post separation (e.g., via centrifugation) of non-

cellular fraction of fluids obtained by puncture of a vein or artery followed by the withdrawal of blood, tissue biopsies, bronchoalveolar lavage, nasal lavage, eye lavage, peritoneal cavity lavage, vaginal lavage, bladder lavage, rectal lavage, fine needle aspiration of spinal fluid, synovial fluid aspiration, and the like. Cell free bodily fluids are obtained post separation (e.g., via centrifugation) of cellular fraction of fluids obtained by puncture of a vein or artery followed by the withdrawal of blood, tissue biopsies, bronchoalveolar lavage, nasal lavage, eye lavage, peritoneal cavity lavage, vaginal lavage, bladder lavage, rectal lavage, fine needle aspiration of spinal fluid, synovial fluid aspiration, and the like.

[0072] In the methods of this invention, cell separation/isolation/purification methods are used to isolate populations of cells from bodily fluid sample, cells, or tissues of a subject. A skilled worker can use any known cell separation/isolation/purification techniques to isolate phagocytic cells, >2n phagocytic cells, diseased cells, control cells, and =2n phagocytic cells from a bodily fluid, or to separate >2n phagocytic cells from =2n phagocytic cells. Exemplary techniques include, but are not limited to, using antibodies, flow cytometry, fluorescence activated cell sorting, filtration, gradient-based centrifugation, elution, microfluidics, magnetic separation technique, fluorescent-magnetic separation technique, nanostructure, quantum dots, high throughput microscope-based platform, or a combination thereof.

[0073] In the methods of this invention, cell separation/isolation/purification methods are used to isolate populations of circulating diseased cells from bodily fluid sample, cells, or tissues of a subject. Circulating diseased cells may be rare or in low quantity in a bodily fluid. Therefore, enrichment techniques (e.g., magnetic enrichment) may be used to enrich circulating diseased cells before the isolation. A skilled worker can use any known cell separation/isolation/purification techniques to isolate circulating diseased cells from a bodily fluid. Exemplary techniques include, but are not limited to, using antibodies, flow cytometry, fluorescence activated cell sorting, filtration, gradient-based centrifugation, elution, microfluidics, magnetic separation technique, fluorescent-magnetic separation technique, nanostructure, quantum dots, high throughput microscope-based platform, micro-fluidic technique,

fiber-optic array–scanning technique, laser-scanning cytometry technique, multiphoton intravital flow cytometry, photoacoustic flowmetry, nanoparticles targeting cell surface antigens, staining circulating diseased cells with detectable secreted products, or a combination thereof. Circulating diseased cells may have different physical properties compared to normal circulating cells, such as difference in size, density, charge, migratory properties, and some properties of specific cell types (e.g., melanocytic granules in circulating melanoma cells). A skilled worker can use any known cell separation/isolation/purification techniques based on such different properties to isolate circulating diseased cells. For example, differences in buoyant density may be used to separate circulating diseased cells (e.g., circulating tumor cells) from normal blood cells through gradient centrifugation. Filtration-based approaches may be used isolate circulating diseased cells (e.g., circulating tumor cells) based on their increased sizes compared to normal circulating cells. Antibody-based isolation approaches may be used to capture circulating diseased cells, which express epithelia cell surface markers that are absent from normal circulating blood cells. For example, conjugation of antibodies against epithelial cell adhesion molecule (EpCAM) to magnetic beads, followed by purification of captured cells through a magnetic field, may be used to enrich circulating tumor cells from the blood of patients with cancers of the breast, prostate, and colon. In certain embodiments, circulating diseased cells (e.g., transcervical cells) may be collected by the RareCellect™ device (Genetic Technologies) or similar devices.

[0074] In certain embodiments, the phagocytic cells, >2n phagocytic cells, diseased cells, control cells, and =2n phagocytic cells are isolated by using a product secreted by the cells. In certain embodiments, the phagocytic cells, >2n phagocytic cells, diseased cells, control cells, and =2n phagocytic cells are isolated by using a cell surface target (e.g., receptor protein) on the surface of the cells. In some embodiments, the cell surface target is a protein that has been engulfed by >2n phagocytic cells. In some embodiments, the cell surface target is expressed by cells on their plasma membranes. In some embodiments, the cell surface target is an exogenous protein that is translocated on the plasma membranes, but not expressed by the cells (e.g., the >2n phagocytic cells). In some embodiments, the cell surface target is a marker of the disease or condition to be detected.

[0075] In certain embodiments, the circulating vesicles are isolated using chromatographic isolation, affinity isolation, or ultracentrifugation.

[0076] In certain aspects of the methods described herein, analytes include nucleic acids, proteins, lipids, carbohydrates, metabolites, or any combinations of these. In certain aspects of the methods described herein, markers include nucleic acids, proteins, lipids, carbohydrates, metabolites, or any combinations of these. As used herein, the term "nucleic acid" is intended to include DNA molecules (e.g., cDNA or genomic DNA), RNA molecules (e.g., mRNA), DNA-RNA hybrids, and analogs of the DNA or RNA generated using nucleotide analogs. The nucleic acid molecule can be a nucleotide, oligonucleotide, double-stranded DNA, single-stranded DNA, multi-stranded DNA, complementary DNA, genomic DNA, non-coding DNA, messenger RNA (mRNAs), microRNA (miRNAs), small nucleolar RNA (snoRNAs), ribosomal RNA (rRNA), transfer RNA (tRNA), small interfering RNA (siRNA), heterogeneous nuclear RNAs (hnRNA), or small hairpin RNA (shRNA). In some embodiments, the nucleic acid is a transrenal nucleic acid. A transrenal nucleic acid is an extracellular nucleic acid that is excreted in the urine, See, e.g., U.S. Patent Publication No. 20100068711 and U.S. Patent Publication No. 20120021404.

[0077] As used herein, the term "amino acid" includes organic compounds containing both a basic amino group and an acidic carboxyl group. Included within this term are natural amino acids (e.g., L-amino acids), modified and unusual amino acids (e.g., D-amino acids and β-amino acids), as well as amino acids which are known to occur biologically in free or combined form but usually do not occur in proteins. Natural protein occurring amino acids include alanine, arginine, asparagine, aspartic acid, cysteine, glutamic acid, glutamine, glycine, histidine, isoleucine, leucine, lysine, methionine, phenylalanine, serine, threonine, tyrosine, tryptophan, proline, and valine. Natural non-protein amino acids include arginosuccinic acid, citrulline, cysteine sulfuric acid, 3,4-dihydroxyphenylalanine, homocysteine, homoserine, ornithine, 3-monoiodotyrosine, 3,5-diiodotryosine, 3, 5, 5-triiodothyronine, and 3,3',5,5'- tetraiodothyronine. Modified or unusual amino acids include D-amino acids, hydroxylysine, 4-hydroxyproline, N-Cbz-protected amino acids, 2,4-diaminobutyric acid, homoarginine, norleucine, N-methylaminobutyric

acid, naphthylalanine, phenylglycine, .alpha.-phenylproline, tert-leucine, 4-aminocyclohexylalanine, N-methyl-norleucine, 3 ,4-dehydroproline, N,N-dimethylaminoglycine, N-methylaminoglycine, 4-aminopiperidine-4-carboxylic acid, 6-aminocaproic acid, trans-4-(aminomethyl)-cyclohexanecarboxylic acid, 2-, 3-, and 4-(aminomethyl)- benzoic acid, 1-aminocyclopentanecarboxylic acid, 1-aminocyclopropanecarboxylic acid, and 2-benzyl-5-aminopentanoic acid.

[0078] As used herein, the term "peptide" includes compounds that consist of two or more amino acids that are linked by means of a peptide bond. Peptides may have a molecular weight of less than 10,000 Daltons, less than 5,000 Daltons, or less than 2,500 Daltons. The term "peptide" also includes compounds containing both peptide and non-peptide components, such as pseudopeptide or peptidomimetic residues or other non-amino acid components. Such compounds containing both peptide and non-peptide components may also be referred to as a "peptide analog."

[0079] As used herein, the term "protein" includes compounds that consist of amino acids arranged in a linear chain and joined together by peptide bonds between the carboxyl and amino groups of adjacent amino acid residues. Proteins used in methods of the invention include, but are not limited to, amino acids, peptides, antibodies, antibody fragments, cytokines, lipoproteins, or glycoproteins.

[0080] As used herein, the term "antibody" includes polyclonal antibodies, monoclonal antibodies (including full length antibodies which have an immunoglobulin Fc region), antibody compositions with polyepitopic specificity, multispecific antibodies (e.g., bispecific antibodies, diabodies, and single-chain molecules, and antibody fragments (*e.g.*, Fab or $F(ab')_2$, and Fv). For the structure and properties of the different classes of antibodies, see e.g., Basic and Clinical Immunology, 8th Edition, Daniel P. Sties, Abba I. Terr and Tristram G. Parsolw (eds), Appleton & Lange, Norwalk, Conn., 1994, page 71 and Chapter 6.

[0081] As used herein, the term "cytokine" refers to a secreted protein or active fragment or mutant thereof that modulates the activity of cells of the immune system. Examples of cytokines include, without limitation, interleukins, interferons,

chemokines, tumor necrosis factors, colony-stimulating factors for immune cell precursors, and the like.

[0082] As used herein, the term "lipoprotein" includes negatively charged compositions that comprise a core of hydrophobic cholesteryl esters and triglyceride surrounded by a surface layer of amphipathic phospholipids with which free cholesterol and apolipoproteins are associated. Lipoproteins may be characterized by their density (e.g. very-low-density lipoprotein (VLDL), low-density lipoprotein (LDL) and high density lipoprotein (HDL)), which is determined by their size, the relative amounts of lipid and protein. Lipoproteins may also be characterized by the presence or absence of particular modifications (e.g. oxidization, acetylation, or glycation).

[0083] As used herein, the term "glycoprotein" includes glycosides which have one or more oligo- or polysaccharides covalently attached to a peptide or protein. Exemplary glycoproteins can include, without limitation, immunoglobulins, members of the major histocompatibility complex, collagens, mucins, glycoprotein IIb/IIIa, glycoprotein-41 (gp41) and glycoprotein-120 (gp12), follicle-stimulating hormone, alpha-fetoprotein, erythropoietin, transferrins, alkaline phosphatase, and lectins.

[0084] As used herein, the term "lipid" includes synthetic or naturally-occurring compounds which are generally amphipathic and biocompatible. Lipids typically comprise a hydrophilic component and a hydrophobic component. Exemplary lipids include, but are not limited to fatty acids, neutral fats, phosphatides, cholesterol, cholesterol esters, triglycerides, glycolipids, glycerolipids, glycerophospholipids, sphingolipids, sterol lipids, prenol lipids, saccharolipids, polyketides, choline glycerophospholipid, ethanolamine glycerophospholipid, phosphatidylinositol, phosphatidylglycerol, phosphatidylserine, lyso-choline glycerophospholipid, lyso-ethanolamine glycerophospholipid, phosphatidic acid, lyso-phosphatidic acid, sphingomyelin, galactosylceramide, glucosylceramide, sulfatide, free fatty acids, prostaglandins, triacylglycerol, diacylglycerol, monoacylglycerol, acyl-CoA, acylcarnitine, oxysterol, ceramide, cardiolipin, sphingoid base-1-phosphate, shingosine, lyso-sphingomyelin,, gangliosides, plasmalogen, sulfatide, ceramide, low

density lipoproteins (LDLs), very low density lipoproteins (VLDLs), high density lipoproteins (HDLs), sphingoid base-1-phosphates or derivatives thereof.

[0085] As used herein, the term "carbohydrate" includes, but is not limited to, compounds that contain oxygen, hydrogen and carbon atoms, typically $(CH_2O)_n$ wherein n is an integer. Exemplary carbohydrates include, but are not limited to, monosaccharides, disaccharides, polysaccharides, or oligosaccharides.

[0086] As used herein, the term "metabolite" includes any molecule used in metabolism. Metabolites can be products, substrates, or intermediates in metabolic processes. Included within this term are primary metabolites, secondary metabolites, organic metabolites, or inorganic metabolites. Metabolites include, without limitation, amino acids, peptides, acylcarnitines, monosaccharides, lipids and phospholipids, prostaglandins, hydroxycicosatetraenoic 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. MoI. 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 cosin 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 (ccff 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	60655170606561	AGTATGGGGATTTTAGCATG	CTTCTGTACTGACCCTTC
rs9389248.T.C	chr6:135282652-135282662	ACACTGACGACATGGTTCTACATAACATCAGATGAGGTCACT	TACGGTAGCAGAGACTTGGTCTGGTTGGCTGTTTATGGTTT
rs1054975.T.C	chr3:9406832-9406842	ACACTGACGACATGGTTCTACACCACTAACCAACTCCTAGAT	TACGGTAGCAGAGACTTGGTCTACAGTGGCTCCAATAGTG
rs6438874.G.A	chr3:124739888-124739898	ACACTGACGACATGGTTCTACACAGACTGACCGTCAAAGA	TACGGTAGCAGAGACTTGGTCTAAGCTTAACAACTCCACTG
rs832068.G.A	chr3:97754143-97754153	ACACTGACGACATGGTTCTACATTTTCCGGGGAAGGTTTT	TACGGTAGCAGAGACTTGGTCTACCTCACTTCTTTGTTTGC
rs840388.T.G	chr5:34018886-34018896	ACACTGACGACATGGTTCTACAACTGGAAAATTACCAAGGC	TACGGTAGCAGAGACTTGGTCTCCGCAAATTCTAAATCTTAAAACA
rs7023954.G.A	chr9:21816754-21816764	ACACTGACGACATGGTTCTACATGATGCCTTAATTTTGGGG	TACGGTAGCAGAGACTTGGTCTCCAGAATTTAAGTTAAATTTAAATTATCCT
rs7808823.T.C	chr7:129989874-129989898	ACACTGACGACATGGTTCTACAGAACAAGGTGGACTTCTG	TACGGTAGCAGAGACTTGGTCTGCTGTAAGCAAGTCCATGA

GenBank Target ID	Target Interval	Left Primer With Tag	Right Primer With Tag
	84		
rs35350755.T.G	chr10:123549687-123549697	ACACTGACGACATGGTTCTACAAAACAAAGCCAATGATTTCC	TACGGTAGCAGAGACTTGGTCTGAAAACTCCAAGTACTGCC
rs3802980.A.G	chr11:5706307-5706317	ACACTGACGACATGGTTCTACAGGTTAGGATTCACTCACCA	TACGGTAGCAGAGACTTGGTCTATTGGCCCATAGTTTATCTTTC
rs2942.G.A	chr6:146755136-146755146	ACACTGACGACATGGTTCTACACACGTGAAGACCAATGAG	TACGGTAGCAGAGACTTGGTCTGTCTTGGTGCTGGTATCT
rs6921.T.A	chr8:103838872-103838882	ACACTGACGACATGGTTCTACACAGACATTTCTTAAAAGTTACCGT	TACGGTAGCAGAGACTTGGTCTCCTTGAACTTCATGGCATA
rs1133181.C.A	chr2:55402268-55402278	ACACTGACGACATGGTTCTACAGAGTTTTCAATCAAATACTTATTCAG	TACGGTAGCAGAGACTTGGTCTGTCGAATTTTGCATAGAAGTAGA
rs3837134.CT.C	chr7:27221061-27221072	ACACTGACGACATGGTTCTACATCTTTCCAAATGTTGCAAGA	TACGGTAGCAGAGACTTGGTCTTTTTCTTTTGCCATTAGTTGAT
rs2414105.A.G	chr15:51868369-51868379	ACACTGACGACATGGTTCTACAAGTCATTTGCCAAAATAACAATA	TACGGTAGCAGAGACTTGGTCTAAATAGTATGTAATAAGATTTCAAAAGT

GenBank Target ID	Target Interval	Left Primer With Tag	Right Primer With Tag
rs7305115.A.G	chr12:72372858-72372868	ACACTGACGACATGGTTCTACACCACTGTACCCAGTACATC	TACGGTAGCAGAGACTTGGTCTGATCATTGACCACTTTATGGT
rs2177142.G.A	chr7:22459320-22459330	ACACTGACGACATGGTTCTACACGGATCCATGTTGACAGA	TACGGTAGCAGAGACTTGGTCTACAGCATTGGATTGCTCA
rs2400478.T.C	chr5:147510862-147510872	ACACTGACGACATGGTTCTACATCTTCTTGAATGCCATAAAGTA	TACGGTAGCAGAGACTTGGTCTGGCTTTAAATCCTTTGGACA
rs6895049.T.G	chr5:52248869-52248879	ACACTGACGACATGGTTCTACAACCAAGAATACATTTAGCATCT	TACGGTAGCAGAGACTTGGTCTCCTGATTTTCCATCCTAGATAA
rs8128.A.C	chr1:115110679-115110689	ACACTGACGACATGGTTCTACAACTTAAACATCAATGGTTTTGG	TACGGTAGCAGAGACTTGGTCTACAATTTAGCAGGTAGAAGAAAA
rs3828977.G.A	chr7:29519494-29519504	ACACTGACGACATGGTTCTACATGACTGGATACCGCTTCT	TACGGTAGCAGAGACTTGGTCTCGGCTTCTGAATAAGGATG
rs12102203.A.G	chr15:51791555-51791565	ACACTGACGACATGGTTCTACATAGATTTAAAGGTCGAATGATCT	TACGGTAGCAGAGACTTGGTCTGAAGCATGCTGTCAAATTT
rs797455	chr12:5023955-	ACACTGACGACATGGTTCTACAC	TACGGTAGCAGAGACTTGGTCTA

GenBank Target ID	Target Interval	Left Primer With Tag	Right Primer With Tag
9.A.T	5023965	ATCTACTAGCATTGGTCC	AGTTTATCTTGGCTTTGGTG
rs2686817.C.A	chr7:47968923-47968933	ACACTGACGACATGGTTCTACATCCCCGAAATCCATCATC	TACGGTAGCAGAGACTTGGTCTTCATGAACCCTGTTCTTAATT
rs1884214.C.T	chr14:37149132-37149142	ACACTGACGACATGGTTCTACATAAATTTTATTTCATGTTTCCTACTTG	TACGGTAGCAGAGACTTGGTCTATAGGTATTTTAAAACATTGTTTAAAACA
rs1053959.C.A	chr9:114359620-114359630	ACACTGACGACATGGTTCTACAATACAGTCCCTTCCAACTC	TACGGTAGCAGAGACTTGGTCTGTTGGCTATCCTACTAATAGTGA
rs3780190.A.G	chr9:139099069-139099079	ACACTGACGACATGGTTCTACAAGTTGTTGAAAGAATCAGCT	TACGGTAGCAGAGACTTGGTCTATGGATGTATTCATAAGACCATG
rs1019503.G.A	chr5:96254813-96254823	ACACTGACGACATGGTTCTACATACAGATAAGTTTTAGAAAGACTCAA	TACGGTAGCAGAGACTTGGTCTATTGCTCCAAAGTGGATTG
rs10883098.C.G	chr10:100218191-100218201	ACACTGACGACATGGTTCTACAGAAGCTTCCTACCTACCA	TACGGTAGCAGAGACTTGGTCTGCCTATTTGAGCTAGAACAAA
rs6979945.G.A	chr7:82386121-82386131	ACACTGACGACATGGTTCTACAAAGCAATTTGATTTATTTCTTTTGG	TACGGTAGCAGAGACTTGGTCTCTAAAGGTGAAGAATTACATGAAAAT

GenBank Target ID	Target Interval	Left Primer With Tag	Right Primer With Tag
	1		
rs147147897.CTA.C	chr16:70192457-70192469	ACACTGACGACATGGTTCTACACTCCTTCTAGACATTTCTTGC	TACGGTAGCAGAGACTTGGTCTCTAACAAGCCAACCAGAAC
rs4692500.A.C	chr4:31147268-31147278	ACACTGACGACATGGTTCTACAATGGGCTATTCCCAACTT	TACGGTAGCAGAGACTTGGTCTAGACTGGTCCATTATCTGAG
rs4727038.G.C	chr7:148979912-148979922	ACACTGACGACATGGTTCTACATGGAAAGTTTTCTACTTTGCT	TACGGTAGCAGAGACTTGGTCTGCAAAGTTCCATTAACAACTAC
rs8692.C.T	chr16:84211479-84211489	ACACTGACGACATGGTTCTACAAGCACCGAAAGCATCATA	TACGGTAGCAGAGACTTGGTCTTGTTTATCATGCCTAAGGGA
rs59508481.C.T	chr3:153839955-153839965	ACACTGACGACATGGTTCTACAAACGGGTTACTAATTACGG	TACGGTAGCAGAGACTTGGTCTCTCCTATGTACCGTCCTG
rs201131957.TTC.T	chr6:87993564-87993576	ACACTGACGACATGGTTCTACAGTGCACCAATTGTAATGAG	TACGGTAGCAGAGACTTGGTCTTAGTCTGAAGAAATCTAACAGGAT
..G.GT	chr15:52102326-52102337	ACACTGACGACATGGTTCTACAATCCTTTAATTAGTTTATGGCTTCA	TACGGTAGCAGAGACTTGGTCTCCTAAATGGGTGAAAGGG

GenBank Target ID	Target Interval	Left Primer With Tag	Right Primer With Tag
rs7096.G.A	chr9:115234607-115234617	ACACTGACGACATGGTTCTACACTGCTAGTGCTGCTTTAC	TACGGTAGCAGAGACTTGGTCTAAAGATCTAATTGGCTTTTATTAGC
..GTC.G	chr12:10311906-10311918	ACACTGACGACATGGTTCTACACCCCACTTGTCCCAAAAT	TACGGTAGCAGAGACTTGGTCTAAAGGTATCAATCCTTGGTTT
rs10486985.G.A	chr7:82386078-82386088	ACACTGACGACATGGTTCTACACACTTCACCCTTAAATAAAGC	TACGGTAGCAGAGACTTGGTCTTGCTGGTCTTACCCTAGT
rs931606.G.A	chr4:79443846-79443856	ACACTGACGACATGGTTCTACATTTCCCCTTCAAATAGGACA	TACGGTAGCAGAGACTTGGTCTAAGTCAAATTCAATTCCTCCTAG
..G.GA	chr11:108345511-108345522	ACACTGACGACATGGTTCTACAACTTGGTGCAGTCTTCTA	TACGGTAGCAGAGACTTGGTCTGAAGAACTTTGAGTCAGCC
rs30169.T.G	chr5:13719018-13719028	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
rs1047481.G.A	chr5:52249472-52249482	ACACTGACGACATGGTTCTACACCAGAAATGTATTTTAAAGAATTTCAG	TACGGTAGCAGAGACTTGGTCTAGTAGATTCCAATGAACCATAATG
rs2293250.G.A	chr3:33434827-33434837	ACACTGACGACATGGTTCTACAAAAACAAATGATTTGATTTACCTGA	TACGGTAGCAGAGACTTGGTCTCACCAAATTAATCAGGTTTACAG
rs10849174.T.C	chr12:5025090-5025100	ACACTGACGACATGGTTCTACAGACTGGTTGATTCTTCTGC	TACGGTAGCAGAGACTTGGTCTCCCAAACTCATAACAACAAC
rs7011928.G.A	chr8:28421981-28421991	ACACTGACGACATGGTTCTACACTCATACAAATAAGAATTGGTAGC	TACGGTAGCAGAGACTTGGTCTATCATCAGGGGCAAGTTA
rs3733690.G.C	chr5:140569523-1405695	ACACTGACGACATGGTTCTACAGAGGCATTTCTTACTAGAATCC	TACGGTAGCAGAGACTTGGTCTGAGGCAATGAATTTAAAGTCAATA

GenBank Target ID	Target Interval	Left Primer With Tag	Right Primer With Tag
	33		
rs1299331.A.G	chr3:106965488-106965498	ACACTGACGACATGGTTCTACAGGTCTTAAGACAGGAAGCTA	TACGGTAGCAGAGACTTGGTCTTCTGATTCCAGTCATAGCTC
rs2278711.C.T	chr2:47085394-47085404	ACACTGACGACATGGTTCTACATGTTCCCTTTCGGATGAG	TACGGTAGCAGAGACTTGGTCTAGTTATTTTGAGCAGATGGTT
rs11692815.G.A	chr2:68415763-68415773	ACACTGACGACATGGTTCTACAACTCGCTGTACTAAAGGAT	TACGGTAGCAGAGACTTGGTCTATGAAATTAAAAGGCTAGGAAAGA
rs116485895.T.C	chr4:69176794-69176804	ACACTGACGACATGGTTCTACACCTTCCTGATGTAATCCTCA	TACGGTAGCAGAGACTTGGTCTCAAGGAAATGGGGTTCTTT
rs2161036.G.T	chr2:186659355-186659365	ACACTGACGACATGGTTCTACAAGTGGAAGAACTTTTGAATAAGT	TACGGTAGCAGAGACTTGGTCTTTGGCAGTTTCATTAAGTGG
rs17229201.T.C	chr2:186656952-186656962	ACACTGACGACATGGTTCTACAAAATACTTAAGTATGTAGTCAAGTTAATTT	TACGGTAGCAGAGACTTGGTCTGCCTGTATCGATAAGTTTCAA
rs2230854.T.C	chr13:103275382-1032753	ACACTGACGACATGGTTCTACAGATGATGGAAACCTGCTC	TACGGTAGCAGAGACTTGGTCTACCACTTTATCACCTCATAAATCT

GenBank Target ID	Target Interval	Left Primer With Tag	Right Primer With Tag
	92		
rs4647554.A.G	chr9:97862697-97862707	ACACTGACGACATGGTTCTACAAAAAGTTCCTGGAATTTTCCA	TACGGTAGCAGAGACTTGGTCTGTTTTGAGGTATGAGCAAAGT
rs704697.T.C	chr11:33724776-33724786	ACACTGACGACATGGTTCTACAGGGGTTAAATAAGAAGTTGATGA	TACGGTAGCAGAGACTTGGTCTTCATAGGGACTTTGCCTC
rs30168.G.A	chr5:13719085-13719095	ACACTGACGACATGGTTCTACAGTTGGCAAATTTAATGGACA	TACGGTAGCAGAGACTTGGTCTATGGACATAATCATAGAAACTGA
rs55938253.A.G	chr2:186661669-186661679	ACACTGACGACATGGTTCTACAAGTTTTGATCAGACCATGAAAG	TACGGTAGCAGAGACTTGGTCTGACAAGACTTGCTTTAATGC
rs10931200.C.A	chr2:186664959-186664969	ACACTGACGACATGGTTCTACAAGAAACAAATCATTTTCTATGCAT	TACGGTAGCAGAGACTTGGTCTGAGTAGAAAATTGCCAGTCT
rs6928.C.G	chr22:22115000-22115010	ACACTGACGACATGGTTCTACACAGATACAAAGCAGTTTCAGA	TACGGTAGCAGAGACTTGGTCTCAGGTTTACCTAAAGCTTGTT
rs2282022.T.C	chr13:41374182-41374192	ACACTGACGACATGGTTCTACAACTAAACAGAATTTGGAAACCA	TACGGTAGCAGAGACTTGGTCTCCTTTGCAATCTTGAAGCTT

GenBank Target ID	Target Interval	Left Primer With Tag	Right Primer With Tag
rs1613780.G.T	chr1:144865846-144865856	ACACTGACGACATGGTTCTACATCTGGAAACATGACTCAGTT	TACGGTAGCAGAGACTTGGTCTCTAACTTCTACAGTCAGGGC
rs7893462.A.G	chr10:28228861-28228871	ACACTGACGACATGGTTCTACAAGAAAGAACGCCAACAAC	TACGGTAGCAGAGACTTGGTCTATCATTGAAAACCTTGTCAAG
rs2275145.G.A	chr14:35242824-35242834	ACACTGACGACATGGTTCTACATTTGTACCTGTCTTTTACTATGA	TACGGTAGCAGAGACTTGGTCTGGAAAACAAGGAAAATGGGA
rs9225.C.T	chr2:47086108-47086118	ACACTGACGACATGGTTCTACACCCCTCCTTAAGCATTGT	TACGGTAGCAGAGACTTGGTCTGATGCTCAAAGTATGATTTCAC
rs1130072.T.C	chr9:14615931-14615941	ACACTGACGACATGGTTCTACAAAGAAATACTTCAGTCCAACA	TACGGTAGCAGAGACTTGGTCTTTATGTATTCTGGATAACTGAAAACA
rs732774.C.T	chr13:52523804-52523814	ACACTGACGACATGGTTCTACAGGATCAATGTCAGTAGATTATTTAAAA	TACGGTAGCAGAGACTTGGTCTTGGATTGTAATCGGTTTTATCG
rs1359300.G.A	chr1:144922579-144922589	ACACTGACGACATGGTTCTACAAGTTTCCTTGAGGTTCTGA	TACGGTAGCAGAGACTTGGTCTGCTGAATTCAGATTCTTCAAG
rs254978	chr5:96230996-	ACACTGACGACATGGTTCTACAT	TACGGTAGCAGAGACTTGGTCTA

GenBank Target ID	Target Interval	Left Primer With Tag	Right Primer With Tag
2.G.T	96231006	ACCTCCTAGTGGTTTGG	CAGCGATAAGTTCCATG
rs9289122.C.G	chr3:118867043-118867053	ACACTGACGACATGGTTCTACAGGGGATTTAAATTTCAGGATG	TACGGTAGCAGAGACTTGGTCTTGAAACTTGGTAAGGATCTTC
rs4766310.G.A	chr12:5022037-5022047	ACACTGACGACATGGTTCTACAGCTACTGACCGATGTTTAAAA	TACGGTAGCAGAGACTTGGTCTGAGTCACTAGGTTTTCTGTTTT
rs4617548.A.G	chr11:16133409-16133419	ACACTGACGACATGGTTCTACACTACTTACCTGGTTTGTATGTTA	TACGGTAGCAGAGACTTGGTCTCGCTCATGATCCCAATTTT
rs10057908.T.G	chr5:102884929-102884939	ACACTGACGACATGGTTCTACAAGTAAAACTGAATTTCAAGATGC	TACGGTAGCAGAGACTTGGTCTATTTAGATACATAATAAAATTCGAGCTA
rs72901775.T.C	chr1:1647810-1647820	ACACTGACGACATGGTTCTACAAAGTCCTCAACTGACCCA	TACGGTAGCAGAGACTTGGTCTAGAAGAACAGGATAAAGCTC
rs2255546.C.T	chr5:96249111-96249121	ACACTGACGACATGGTTCTACATGGAAGGAAAGGTTATCAAGA	TACGGTAGCAGAGACTTGGTCTAGAAGATGGGTCCAATTTTC
rs2548538.T.A	chr5:96232138-96232148	ACACTGACGACATGGTTCTACACTCCCTGTTTAGGATGACTA	TACGGTAGCAGAGACTTGGTCTATTTCCTGTATTTGAGTCGG
rs177083.	chr2:324	ACACTGACGACATGGTTCTACAG	TACGGTAGCAGAGACTTGGTCTG

GenBank Target ID	Target Interval	Left Primer With Tag	Right Primer With Tag
C.T	47714-32447724	TGACCTGCTAAAGAAATACA	ATAAAGGCCAAGGAACAG
rs6876143.A.G	chr5:17380266-17380276	ACACTGACGACATGGTTCTACATGAAAACTGTCTTTCATTAGCT	TACGGTAGCAGAGACTTGGTCTAATACCATGTCTGCCTGT
rs9841174.T.C	chr3:167184874-167184884	ACACTGACGACATGGTTCTACACATGAATACAATGGCAAATAATAAT	TACGGTAGCAGAGACTTGGTCTGTGGATTTTCAAGATGCAAAG
rs116571982.G.A	chr6:31023144-31023154	ACACTGACGACATGGTTCTACAAAGACTTGAACCCTATTAGAAAA	TACGGTAGCAGAGACTTGGTCTGAATACATGTAGCTTACATGGC
rs4898.T.C	chr7:82386202-82386212	ACACTGACGACATGGTTCTACATTCTTCACCTTTAGTCAGGTTA	TACGGTAGCAGAGACTTGGTCTATCATCTAAACCAATGGCAAG
rs202026726.T.TA	chr3:86990606-86990617	ACACTGACGACATGGTTCTACAAAAGTATGTTGCATTTTAAAAGAATTAT	TACGGTAGCAGAGACTTGGTCTCACCTTTTAGAAAAGGATAAGTAAT
rs11537748.A.G	chr15:31668690-31668700	ACACTGACGACATGGTTCTACATTTTGCTGACTGTTTCCATA	TACGGTAGCAGAGACTTGGTCTGGGGTAATCCTGATTTCAAAA
rs14026.G.A	chr2:55404790-55404800	ACACTGACGACATGGTTCTACAAGTAGCTTAATGCCAACTTT	TACGGTAGCAGAGACTTGGTCTAGAAACAACCATCTTTATCTTTTG

GenBank Target ID	Target Interval	Left Primer With Tag	Right Primer With Tag
	0		
rs1780138.A.G	chr10:38501650-38501660	ACACTGACGACATGGTTCTACAAAAGTGAAGACTCCTTCTACC	TACGGTAGCAGAGACTTGGTCTAGTTAAACTATTGTTTTAAAAGCCTT
rs3800951.T.G	chr7:99758132-99758142	ACACTGACGACATGGTTCTACACATCGAAGTACTCAGCCA	TACGGTAGCAGAGACTTGGTCTTCGATTTGGGGTCTTCAT
rs1230382.T.A	chr5:96254350-96254360	ACACTGACGACATGGTTCTACATACCATTTGGTTTAAGCCTTAC	TACGGTAGCAGAGACTTGGTCTGGCAGATTCGACTTCATTT
rs201009141.CCT.C	chr7:27221060-27221072	ACACTGACGACATGGTTCTACATCTTTCCAAATGTTGCAAGA	TACGGTAGCAGAGACTTGGTCTTTTTTCTTTTGCCATTAGTTGAT
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 °C for 10 minutes
16 cycles of the
three steps
below
95 °C for 15 seconds
60 °C for 30 seconds
72 °C for 1 minute
72 °C for 3 minutes
4 °C 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.

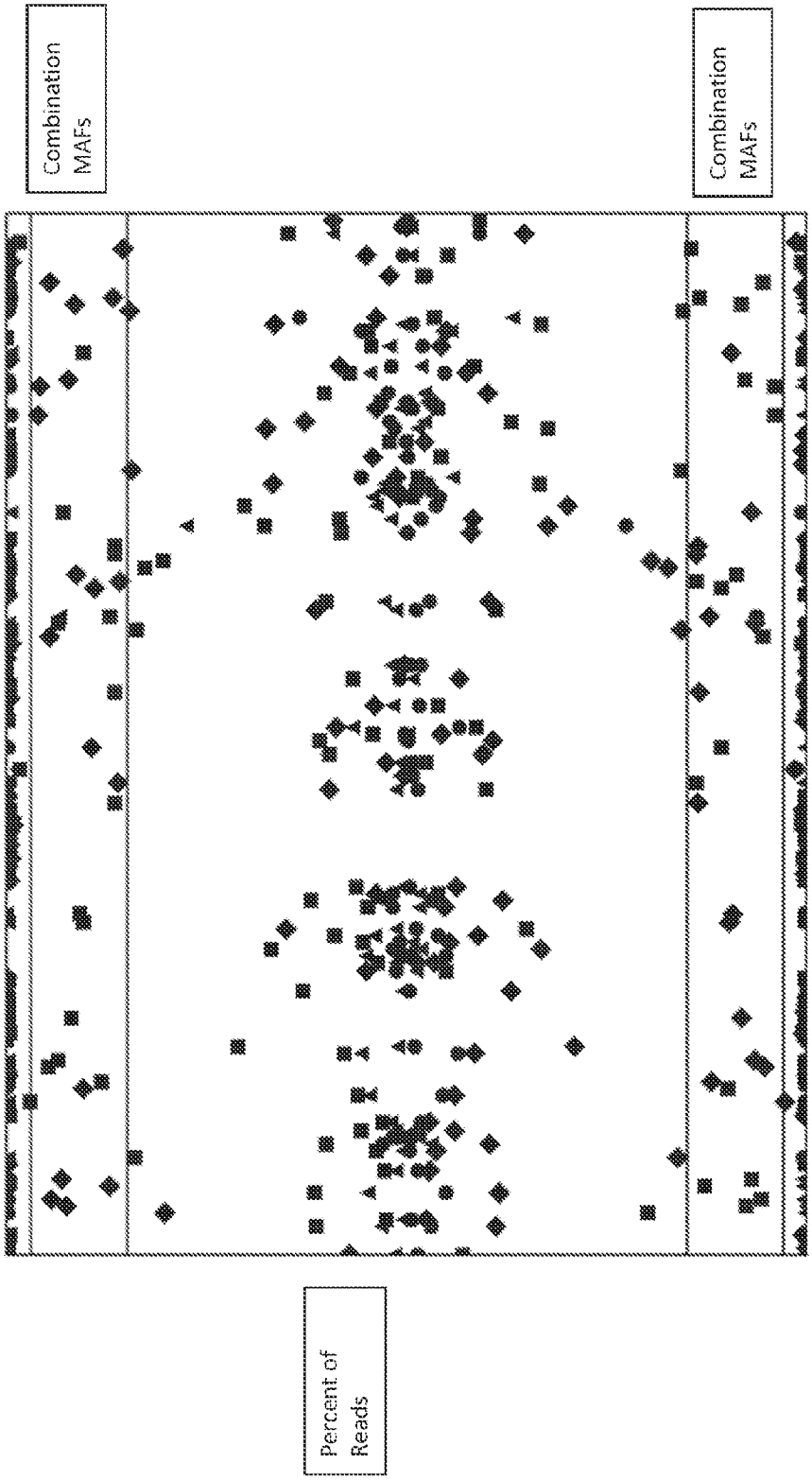


Figure 1

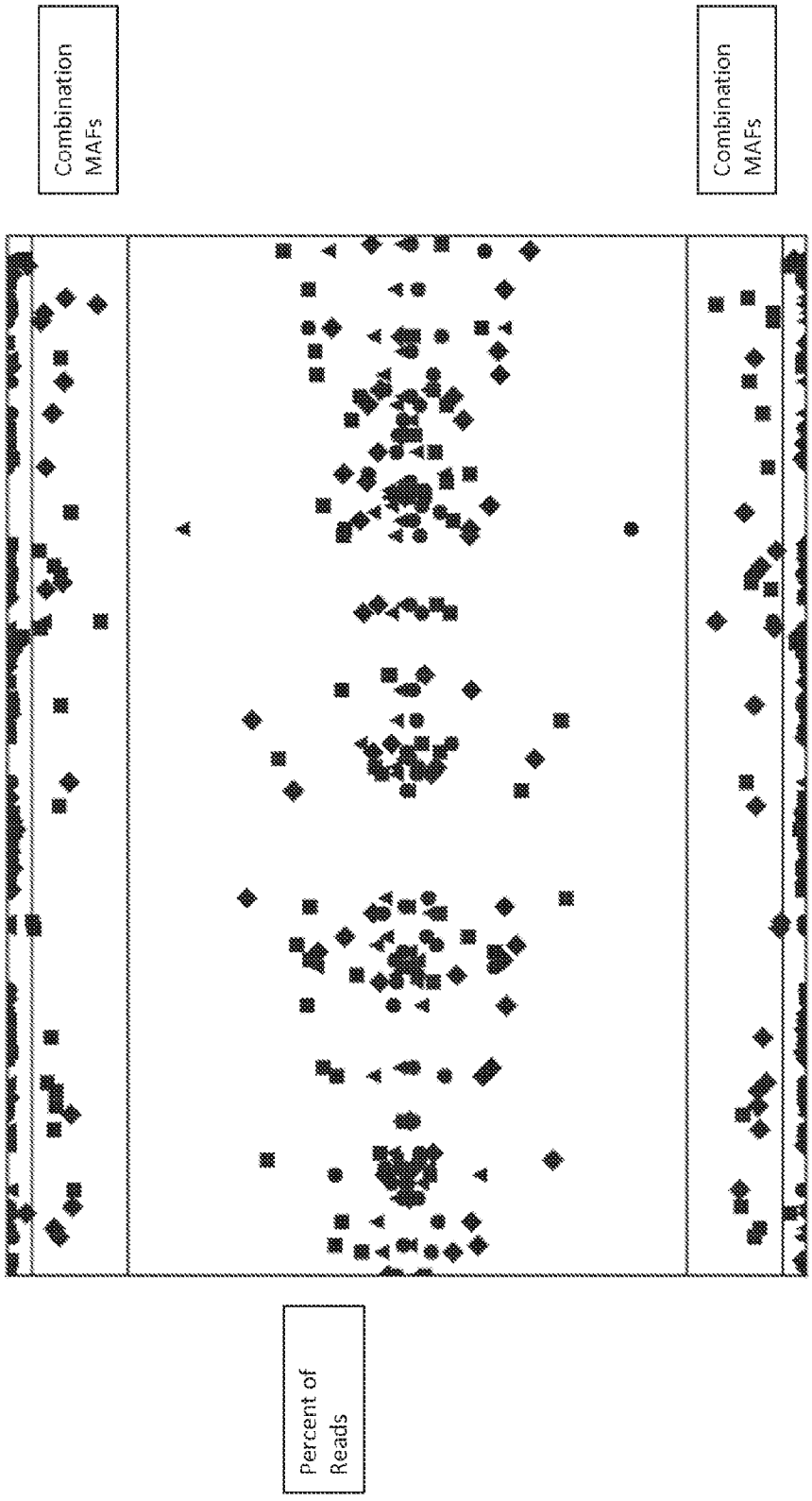


Figure 2

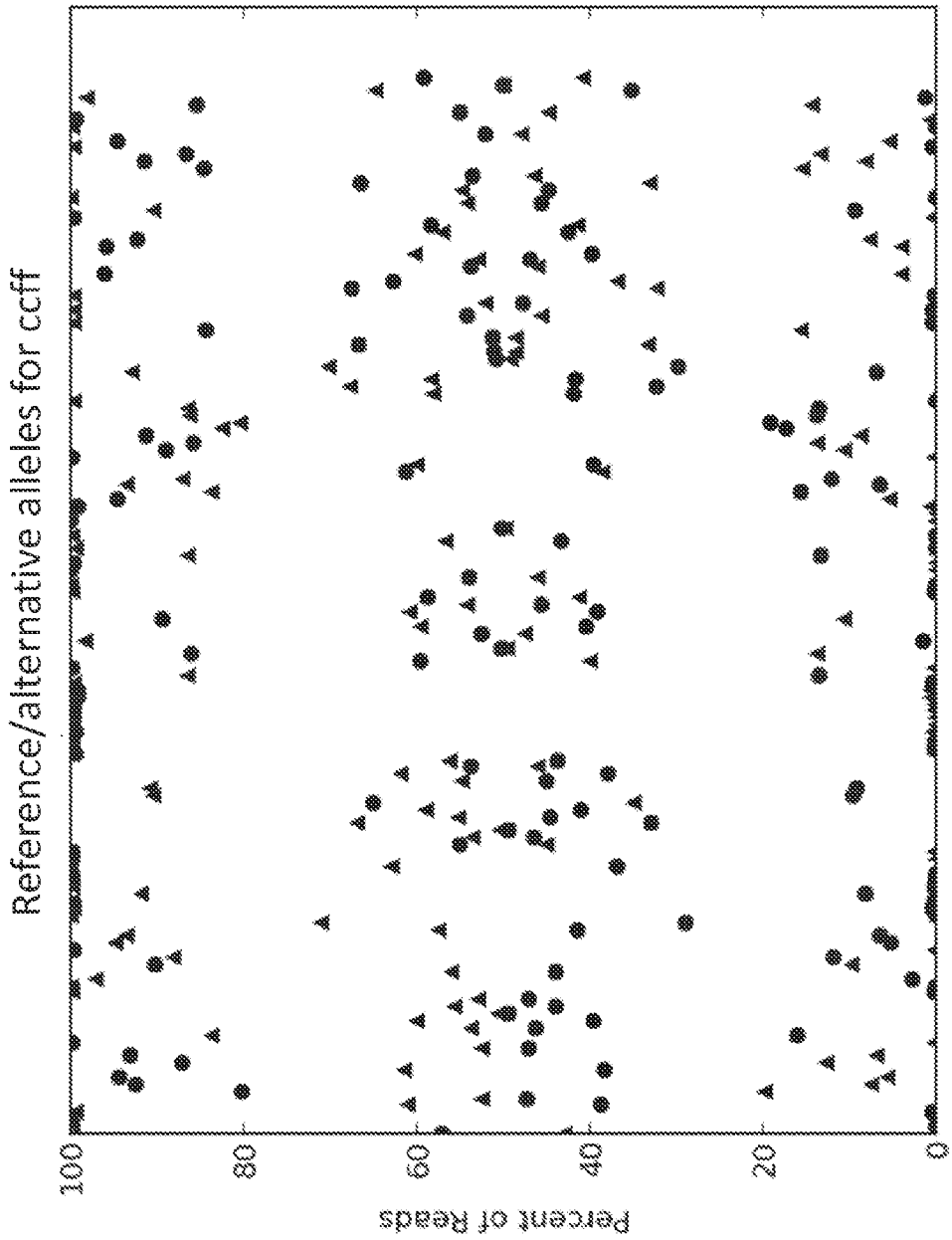

Reference/alternative alleles for ccff
Percent of Reads
100
80
60
40
20
0

Figure 3

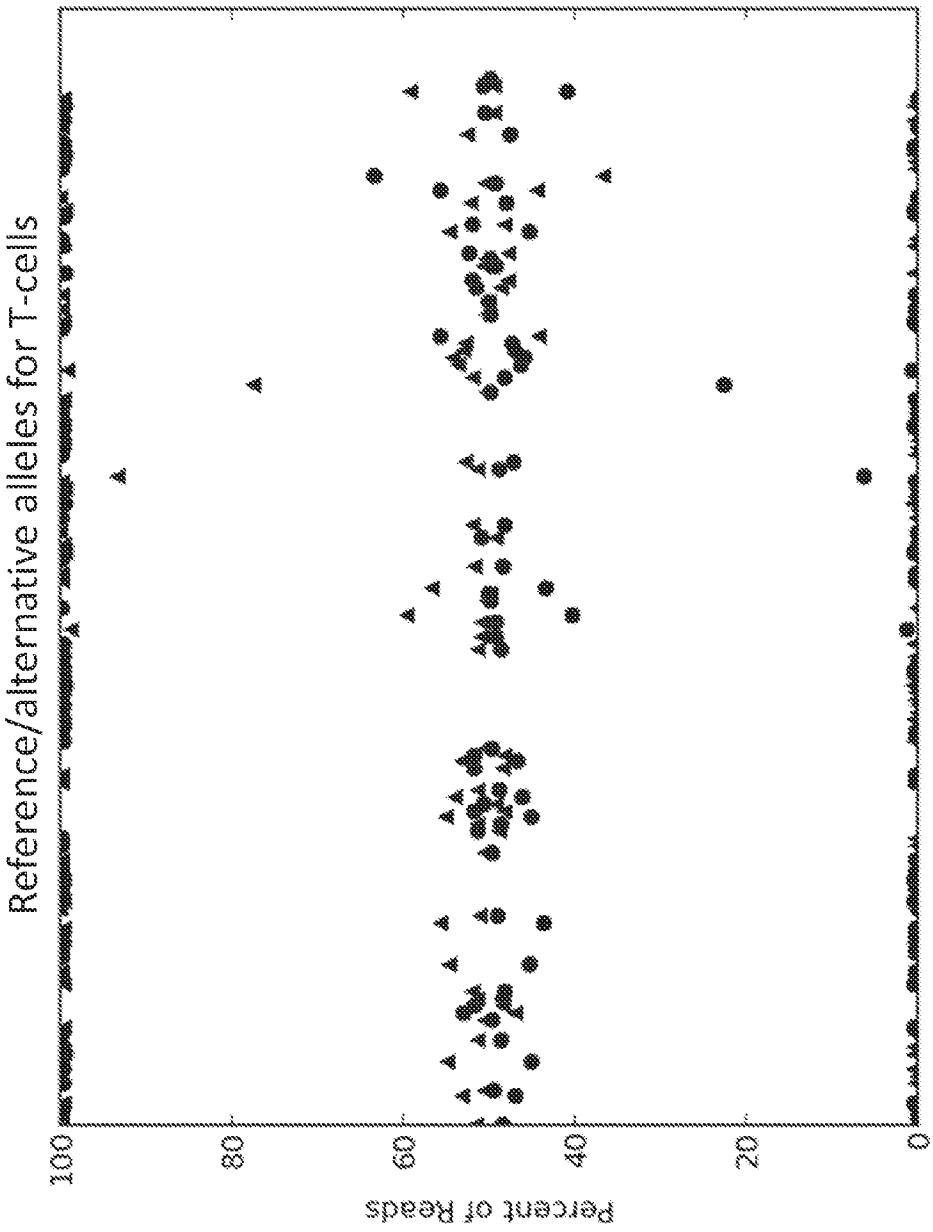

Reference/alternative alleles for T-cells
Percent of Reads
100
80
60
40
20
0

Figure 4

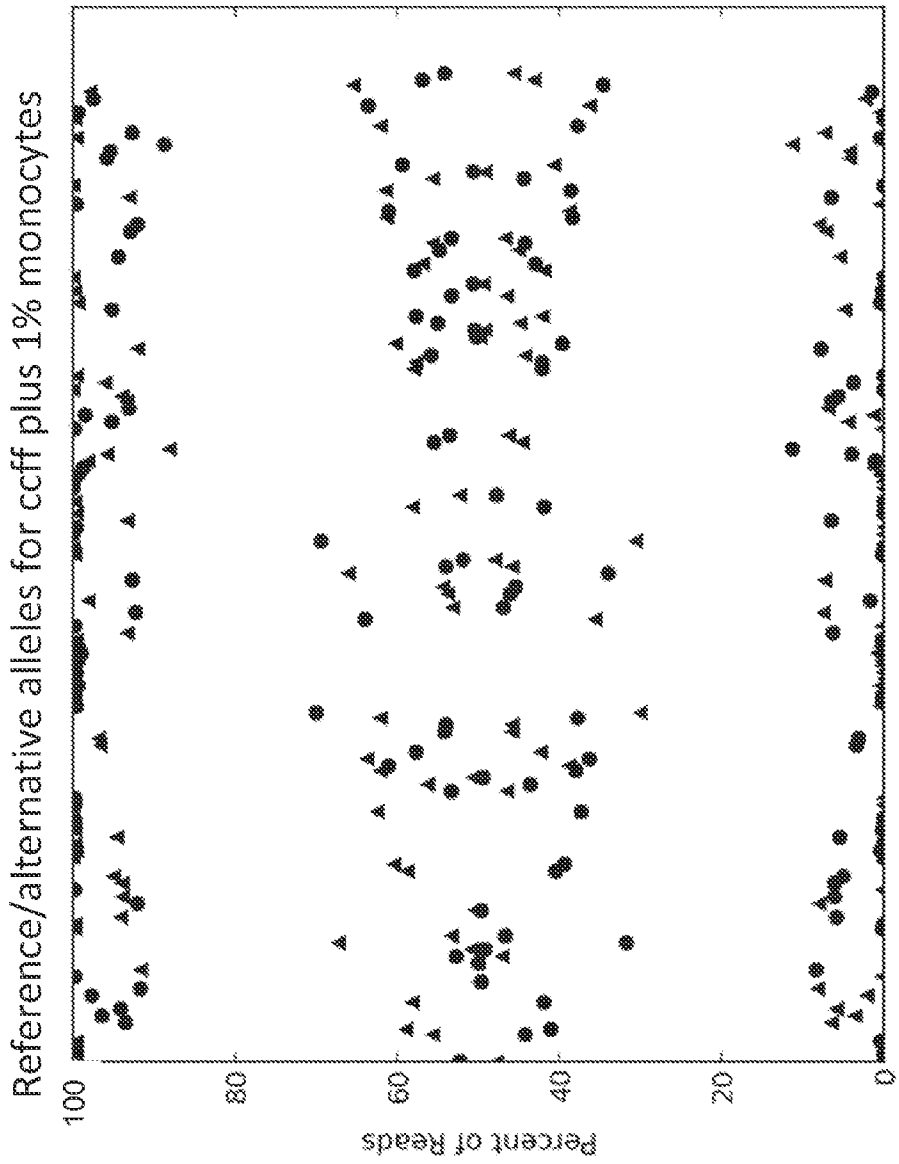

Reference/alternative alleles for ccff plus 1% monocytes
Percent of Reads
100
80
60
40
20
0

Figure 5

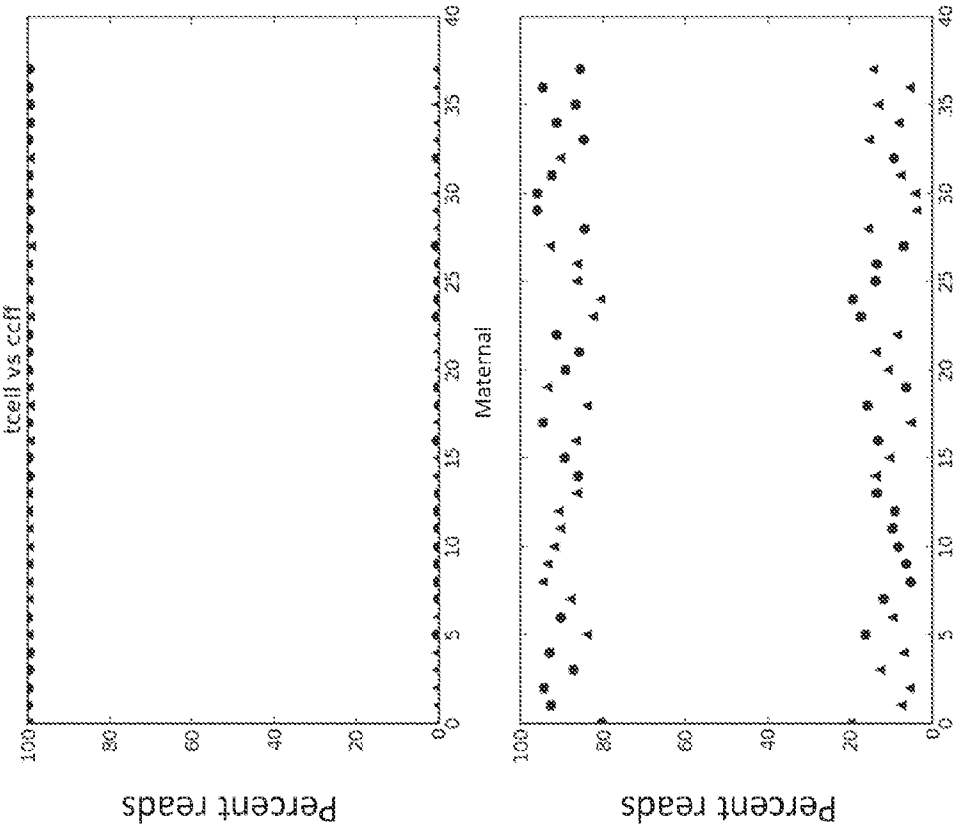

tcell vs cdf
Percent reads
Maternal
Percent reads
Combination

Figure 6

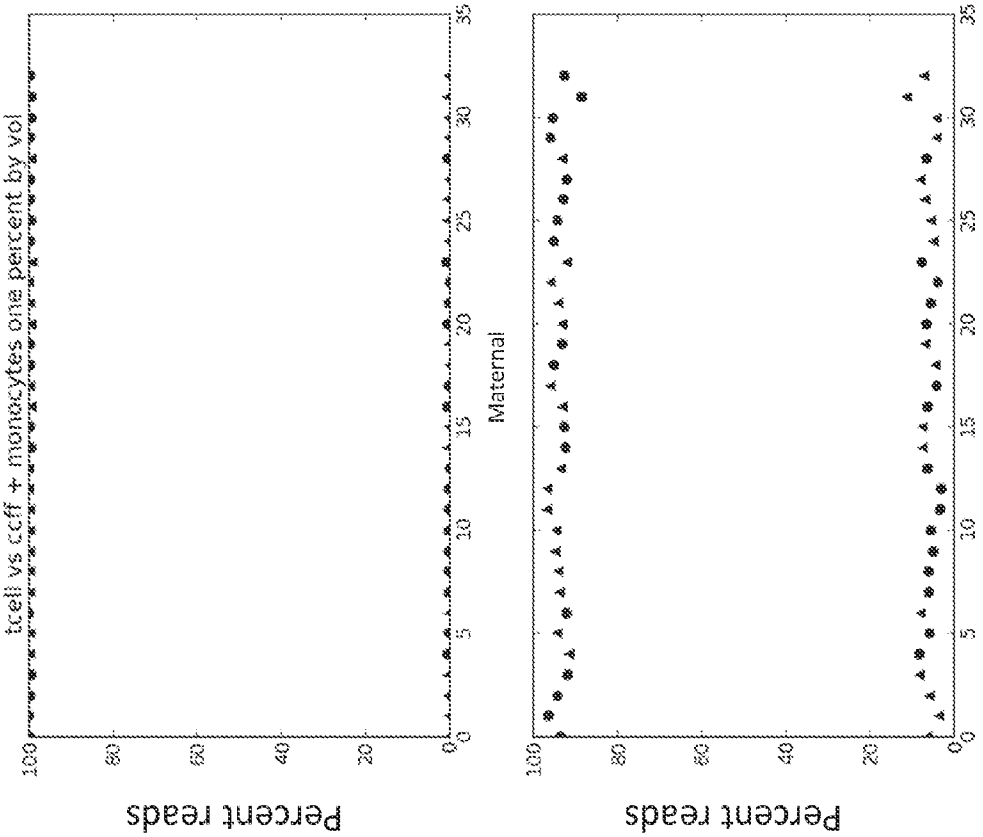

tcell vs cdff + monocytes one percent by vol
Percent reads
Maternal
Percent reads
Combination

Figure 7

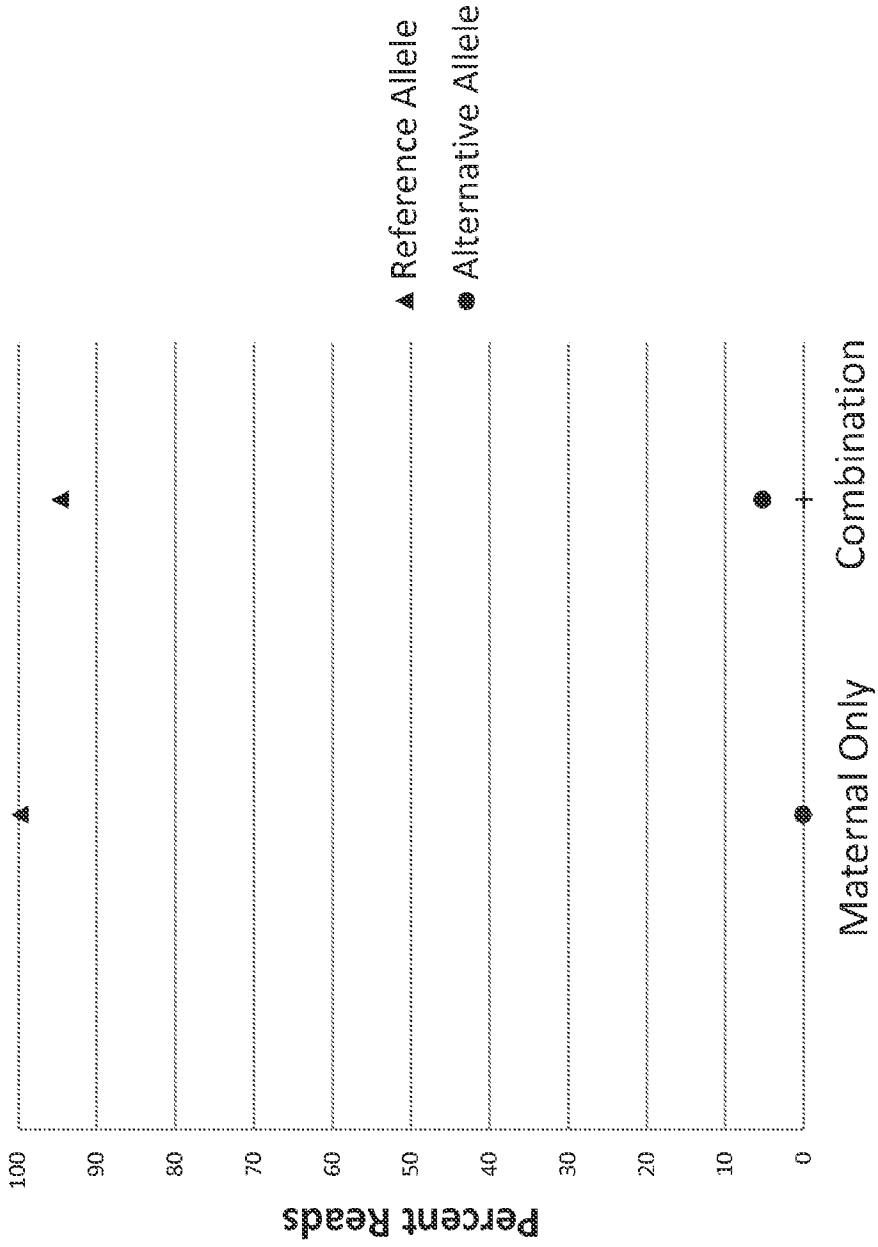

Reference Allele
Alternative Allele
Percent Reads
100
90
80
70
60
50
40
30
20
10
0
Maternal Only
Combination

Figure 8

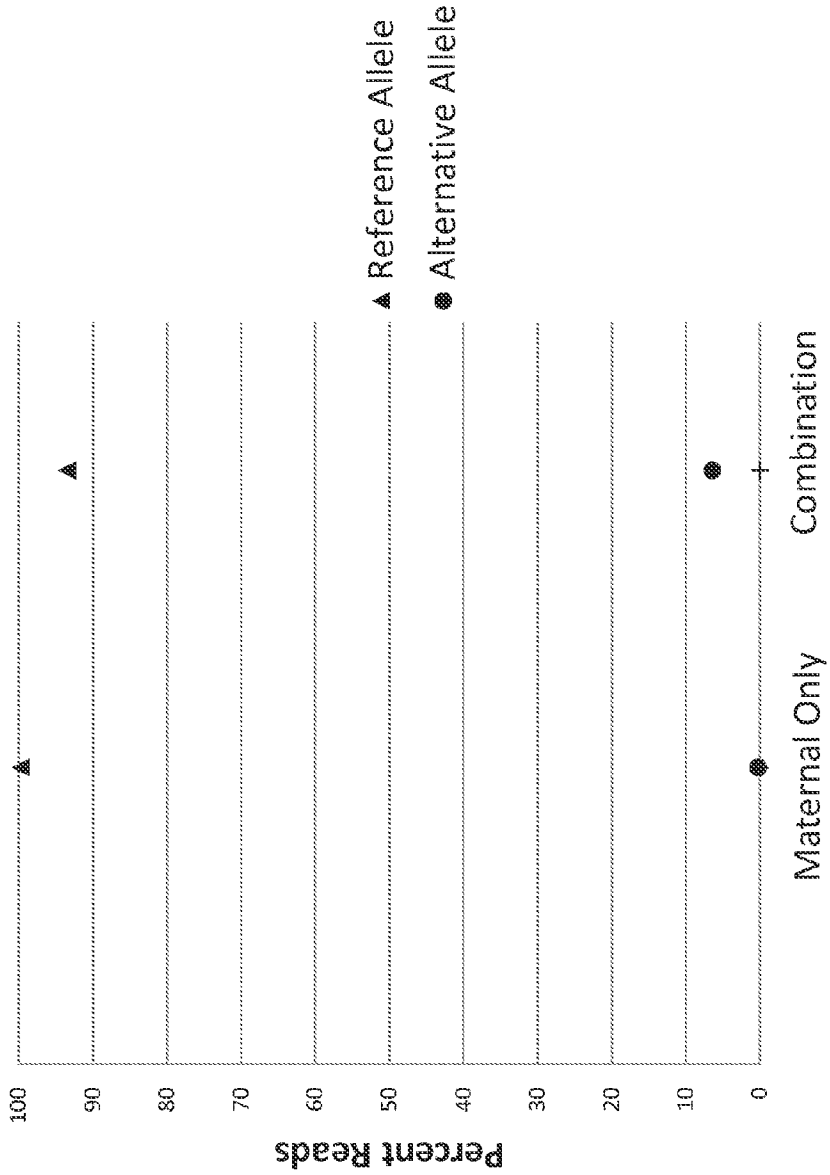

Reference Allele
Alternative Allele
Percent Reads
100
90
80
70
60
50
40
30
20
10
0
Maternal Only
Combination

Figure 9

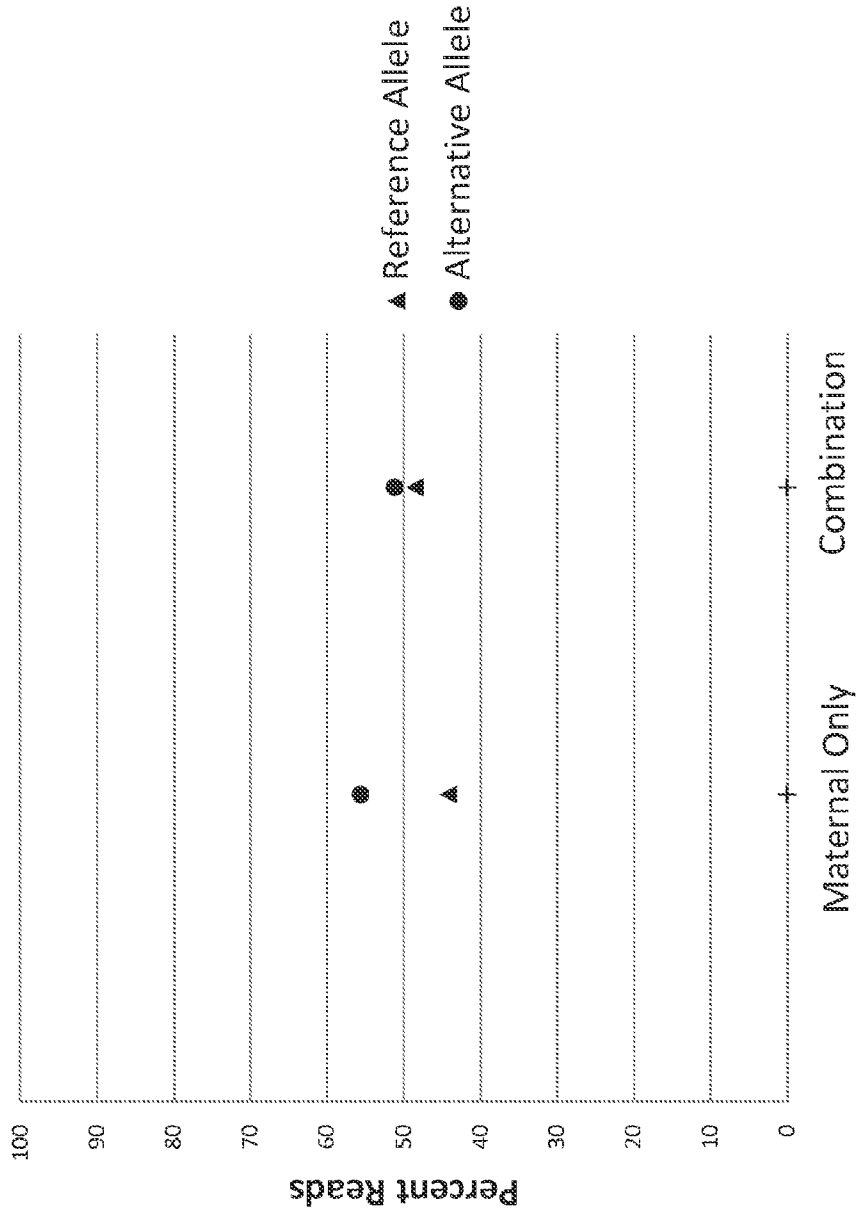

Reference Allele
Alternative Allele
Percent Reads
100
90
80
70
60
50
40
30
20
10
0
Maternal Only
Combination

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.

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); G16B30/00 (EP, 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)**;
- C12Q1/6869, C12Q2535/122, C12Q2537/165 **(EP)**

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 葛会娟 李旭超 易赏 汪建 王俊 杨焕明 张秀清
(74)专利代理机构 北京清亦华知识产权代理事务所(普通合伙) 11201
代理人 李志东
(51)Int.Cl.
C12Q 1/68(2006.01)

(54)发明名称

确定胎儿基因组中预定区域碱基信息的方法、系统和计算机可读介质

(57)摘要

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

CN 104053789 A

Espacenet

Bibliographic data: CN104053789 (B) — 2016-02-10

Method, System And Computer Readable Medium For Determining Base Information In Predetermined Area Of Fetus Genome

Inventor(s): CHEN SHENGPEI, ; GE HUIJUAN, ; LI XUCHAO, ; YI SHANG, ; WANG JIAN, ; WANG JUN, ; YANG HUANMING, ; ZHANG XIUQIN

Applicant(s): BGI HEALTH SERVICE CO LTD

Classification: - international: *C12Q1/68*
- cooperative: default C12Q1/6809 (CN); C12Q1/6858 (EP, KR, US); C12Q1/6869 (EP, KR, US); C12Q1/6883 (US); G16B20/00 (EP, KR, US); G16B30/00 (EP, KR, US); G16B40/00 (KR); G16H50/20 (EP, KR, US); G16H50/30 (EP, KR, US); C12Q2600/118 (US); C12Q2600/156 (US)

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

less

Application number: CN20128067404 20120514 Global Dossier

Priority number(s): WO2012CN75478 20120514

Also published as: CN104053789 (A) AU2012380221 (A1) AU2012380221 (B2) EP2851431 (A1) EP2851431 (A4) EP2851431 (B1) ES2656023 (T3) HK1196401 (A1) JP2015525062 (A) JP6045686 (B2) KR101770460 (B1) KR20140146193 (A) PL2851431 (T3) RU2014150655 (A) RU2597981 (C2) SG11201407515R (A) US2015094210 (A1) US2018320235 (A1) WO2013170429 (A1) less

Abstract of CN104053789 (A)

Provided are a method, system and computer readable medium for determining the base information in a predetermined area of a fetus genome, the method comprising the following steps: constructing a sequence library for the DNA samples of the fetus genome; sequencing the sequence library to obtain the sequencing result of the fetus,

the sequencing result of the fetus comprised of a plurality of sequencing data; and based on the sequencing result of the fetus, determining the base information in the predetermined area according to the hidden Markov model in conjunction with the genetic information of an individual related hereditarily to the fetus.

(19)中华人民共和国国家知识产权局

(12)发明专利

(10)授权公告号 CN 104053789 B
(45)授权公告日 2016.02.10

(21)申请号 201280067404.1

(22)申请日 2012.05.14

(85)PCT国际申请进入国家阶段日
2014.07.17

(86)PCT国际申请的申请数据
PCT/CN2012/075478 2012.05.14

(87)PCT国际申请的公布数据
WO2013/170429 ZH 2013.11.21

(73)专利权人 深圳华大基因股份有限公司
地址 518083 广东省深圳市盐田区洪安三街21号华大综合园7栋7层-14层

(72)发明人 陈盛培 葛会娟 李旭超 易赏 汪建 王俊 杨焕明 张秀清

(74)专利代理机构 北京清亦华知识产权代理事务所(普通合伙) 11201
代理人 李志东

(51)Int.Cl.
C12Q 1/68(2006.01)

(56)对比文件
CN 102317473 A,2012.01.11, 说明书第77-161段.
CN 102127818 A,2011.07.20, 说明书第8-37段.

审查员 蔡放

权利要求书3页 说明书14页 附图2页

(54)发明名称

确定胎儿基因组中预定区域碱基信息的方法、系统和计算机可读介质

(57)摘要

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

CN 104053789 B

1. 一种确定胎儿基因组中预定区域碱基信息的方法，其特征在于，包括下列步骤：

针对胎儿基因组 DNA 样本，构建测序文库；

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

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

将所述胎儿的单倍型作为隐含状态，将所述胎儿的测序结果当作观察序列，借助先验数据推算出重组概率、观察序列概率分布和初始状态概率分布，按照惠特比算法，推断出最可能的所述胎儿的单体型。

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

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

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

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

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

7. 根据权利要求 1 所述的方法，其特征在于，在所述惠特比算法中，采用 0.25 作为所述初始状态概率分布，采用 re/N 作为所述重组概率，其中 re = 25 ～ 30，N 为所述预定区域的长度，

采用

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

作为重组转移矩阵，其中，p_r = re/N。

8. 根据权利要求 7 所述的方法，其特征在于，re = 25。

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

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

其中，

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

。

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

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

12. 一种用于确定胎儿基因组中预定区域碱基信息的系统，其特征在于，包括：

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

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

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

将所述胎儿的单倍型作为隐含状态，将所述胎儿的测序结果当作观察序列，借助先验数据推算出重组概率、观察序列概率分布和初始状态概率分布，按照惠特比算法，推断出最可能的所述胎儿的单体型。

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

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

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

16. 根据权利要求 12 所述的系统，其特征在于，所述惠特比算法采用 0.25 作为所述初始状态概率分布，采用 re/N 作为所述重组概率，其中 re = 25 ～ 30，N 为所述预定区域的长度，

采用

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

作为重组转移矩阵，其中，p_r = re/N。

17. 根据权利要求 16 所述的系统，其特征在于，re = 25。

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

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

其中，

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

。

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

[0001] 优先权信息

[0002] 无

技术领域

[0003] 本发明涉及确定胎儿基因组中预定区域碱基信息的方法、系统和计算机可读介质。

背景技术

[0004] 遗传性疾病是由于遗传物质发生改变而造成的疾病，具有先天性、家族性、终身性和遗传性的特点。遗传性疾病可分为 3 个大类：单基因遗传病、多基因遗传病及染色体异常。其中单基因病多由于单个致病基因的显性或隐性遗传所致基因功能异常；而多基因遗传病则是由多个基因变化影响所致的疾病，会在一定程度上受到外界环境因素的影响；染色体异常包括数目异常和结构异常，最为多见的是由于第 21 号染色体三体所致的唐氏综合症，患儿表现为先天愚型和肢体形状异常等其他先天性特征。由于目前对遗传性疾病尚无有效的治疗方式，只能针对性地进行支持治疗或者药物缓解，费用昂贵，给社会和家庭带来沉重经济和精神负担。因此，在孩子出生前就对孩子的患病状态进行检测，做好预防工作，以达到优生优育的目的，是十分必要的。

[0005] 然而，目前的相关检测手段仍有待改进。

发明内容

[0006] 本发明旨在至少解决现有技术中存在的技术问题之一。

[0007] 在本发明的一个方面，本发明提出了一种确定胎儿基因组中预定区域碱基信息的方法。根据本发明的实施例，该方法包括下列步骤：针对胎儿基因组 DNA 样本，构建测序文库；对所述测序文库进行测序，以便获得胎儿的测序结果，所述胎儿的测序结果由多个测序数据构成；基于所述胎儿的测序结果，结合胎儿遗传相关个体的遗传信息，根据隐马尔可夫模型，确定所述预定区域的碱基信息。子代的基因组形成过程，相当于亲代基因组的一次随机重组（即连锁互换单倍体型重组，以及配子的随机组合）。对于孕期血浆，假若我们将胎儿的单倍型（父母单倍型的重组型）作为隐含状态 (hidden states)，可以将血浆的测序数据当做观察序列 (observations)，借助先验数据推算出状态转移概率 (transition probabilities)、观察序列概率分布 (observation symbol probabilities) 和初始状态概率分布 (initial state distribution)，我们则可以通过诸如惠特比算法 (Viterbi algorithm) 根据隐马尔可夫模型推断出最可能的胎儿单倍型组合，从而获得更多胎儿的信息。因而，根据本发明的实施例，借助隐马尔可夫模型，例如可以通过利用惠特比算法 (Viterbi algorithm)，参考胎儿遗传相关个体的遗传信息，可以确定胎儿基因组中特定区域的核酸序列，由此，可以有效地对胎儿基因组的遗传信息进行产前检测。

[0008] 在本发明的又一方面，本发明提出了一种用于确定胎儿基因组中预定区域碱基信息的系统。根据本发明的实施例，该系统包括：文库构建装置，所述文库构建装置适于针对胎儿基因组 DNA 样本，构建测序文库；测序装置，所述测序装置与所述文库构建装置相连，并且适于对所述测序文库进行测序，以便以便获得胎儿的测序结果，所述胎儿的测序结果由多个测序数据构成；分析装置，基于所述胎儿的测序结果，结合胎儿遗传相关个体的遗传信息，根据隐马尔可夫模型，确定所述预定区域的碱基信息。利用该系统，能够有效地实施前面所述的确定胎儿基因组中预定区域碱基信息的方法，可以借助隐马尔可夫模型，例如可以通过利用惠特比算法 (Viterbi algorithm)，参考胎儿遗传相关个体的遗传信息，可以确定胎儿基因组中特定区域的核酸序列，由此，可以有效地对胎儿基因组的遗传信息进行产前检测，从而可以有效地对胎儿基因组的遗传信息进行产前确定。

[0009] 在本发明的另一方面，本发明还提出了一种计算机可读介质。根据本发明的实施例，该计算机可读介质上存储有指令，所述指令适于被处理器执行以便基于胎儿的测序结果，结合胎儿遗传相关个体的遗传信息，根据隐马尔可夫模型，确定预定区域的碱基信息。利用本发明的计算机可读介质，能够有效地被处理器执行其存储的指令，以便借助隐马尔可夫模型，例如可以通过利用惠特比算法 (Viterbi algorithm)，基于胎儿的测序结果，参考胎儿遗传相关个体的遗传信息，可以确定胎儿基因组中特定区域的核酸序列，由此，可以有效地对胎儿基因组的遗传信息进行产前检测。

[0010] 本发明的附加方面和优点将在下面的描述中部分给出，部分将从下面的描述中变得明显，或通过本发明的实践了解到。

附图说明

[0011] 本发明的上述和 / 或附加的方面和优点从结合下面附图对实施例的描述中将变得明显和容易理解，其中：

[0012] 图 1 为根据本发明一个实施例的利用隐马尔可夫模型进行分析的流程示意图；以及

[0013] 图 2 为根据本发明的一个实施例的用于确定胎儿基因组中预定区域核酸序列的系统的结构示意图。

具体实施方式

[0014] 下面详细描述本发明的实施例，所述实施例的示例在附图中示出，其中自始至终相同或类似的标号表示相同或类似的元件或具有相同或类似功能的元件。下面通过参考附图描述的实施例是示例性的，仅用于解释本发明，而不能理解为对本发明的限制。

[0015] 需要说明的是，术语“第一”、“第二”仅用于描述目的，而不能理解为指示或暗示相对重要性或者隐含指明所指示的技术特征的数量。由此，限定有“第一”、“第二”的特征可以明示或者隐含地包括一个或者更多个该特征。进一步地，在本发明的描述中，除非另有说明，“多个”的含义是两个或两个以上。

[0016] 确定胎儿基因组中预定区域碱基信息的方法

[0017] 在本发明的第一方面，本发明提出了一种确定胎儿基因组中预定区域碱基信息的方法。根据本发明的实施例，该方法包括下列步骤：

[0018] 首先，针对胎儿基因组 DNA 样本，构建测序文库。根据本发明的实施例，胎儿基因组 DNA 样本的来源并不受特别限制。根据本发明的一些实施例，可以采用任何含有胎儿核酸的孕妇样本。例如，根据本发明的实施例，可以采用的孕妇样本为孕妇乳汁、尿液和外周血。其中，优选孕妇外周血。采用孕妇外周血作为胎儿基因组 DNA 样本的来源，可以有效地实现无创取样方式获得胎儿基因组 DNA，从而可以在不影响胎儿正常发育的前提下，对胎儿的基因组进行有效监测。关于针对核酸样本，构建测序文库的方法和流程，本领域技术人员可以根据不同的测序技术进行适当选择，关于流程的细节，可以参见测序仪器的厂商例如 Illumina 公司所提供的规程，例如参见 Illumina 公司 Multiplexing Sample Preparation Guide (Part#1005361 ;Feb2010) 或 Paired-End SamplePrep Guide(Part#1005063 ; Feb2010)，通过参照将其并入本文。根据本发明的实施例，从生物样本提取核酸样本的方法和设备，也不受特别限制，可以采用商品化的核酸提取试剂盒进行。

[0019] 在构建测序文库后，将测序文库应用于测序仪器，对测序文库进行测序，并获得相应的测序结果，该测序结果是由多个测序数据构成的。根据本发明的实施例，可以用于进行测序的方法和设备并不受特别限制，包括但不限于双脱氧链终止法；优选高通量的测序方法，由此，能够利用这些测序装置的高通量、深度测序的特点，进一步提高测序效率。从而，能够提高后续对测序数据进行分析，尤其是统计检验分析时的精确性和准确度。所述高通量的测序方法包括但不限于第二代测序技术或者是单分子测序技术。所述第二代测序平台 (MetzkerML. Sequencing technologies-the next generation. Nat Rev Genet. 2010Jan ;11(1) :31-46) 包括但不限于 Illumina-Solexa(GA™, HiSeq2000™等)、ABI-Solid 和 Roche-454(焦磷酸测序)测序平台；单分子测序平台(技术)包括但不限于 Helicos 公司的真实单分子测序技术 (True Single Molecule DNA sequencing)，Pacific Biosciences 公司单分子实时测序 (single molecule real-time(SMRT™))，以及 Oxford Nanopore Technologies 公司的纳米孔测序技术等 (Rusk, Nicole(2009-04-01). Cheap Third-Generation Sequencing. Nature Methods6(4) :244-245)。随着测序技术的不断进化，本领域技术人员能够理解的是还可以采用其他的测序方法和装置进行全基因组测序。根据本发明的具体示例，可以利用选自 Illumina-Solexa、ABI-SOLiD、Roche-454 和单分子测序装置的至少一种对所述全基因组测序文库进行测序。

[0020] 任选地，在得到测序结果之后，可以将所述测序结果与参照序列进行比对，以便确定与所述预定区域对应的测序数据。在本文中所使用的术语“预定区域”应作广义理解，是指任何包含可能发生预定事件位点的核酸分子的区域。对于 SNP 分析而言，可以是指包含 SNP 位点的区域。对于分析染色体非整倍性，则预定区域指的是所要分析的染色体的全长或者部分，即选择所有来自该染色体的测序数据。从测序结果中选择来自相应区域的测序数据的方法可以不受特别限制。根据本发明的实施例，可以通过将所得到的所有测序数据与已知的核酸参照序列进行比对，从而得到来自于预定区域的测序数据。另外，根据本发明的实施例，预定区域也可以是基因组上不连续的多个分散点。根据本发明的实施例，可以使用的参照序列的类型并不受特别限制，可以为任何含有感兴趣区域的已知序列。根据本发明的实施例，可以采用已知的人类参考基因组作为参照序列。例如，根据本发明的实施例，采用的人类参考基因组为 NCBI36. 3，HG18。另外，根据本发明的实施例，进行比对的方法并不受特别限制。根据本发明的具体实施例，可以采用 SOAP 进行比对。

[0021] 接下来，基于与预定区域对应的测序数据，确定预定区域中的部分核酸序列；以及基于所确定的预定区域中的部分核酸序列，按照惠特比算法，确定预定区域的其他核酸序列，以便获得预定区域的核酸序列。根据本发明的实施例，可以通过基于所述胎儿的测序结果，结合胎儿遗传相关个体的遗传信息，根据隐马尔可夫模型，确定所述预定区域的碱基信息。根据本发明的实施例，可以利用惠特比算法 (Viterbi algorithm)，借助隐马尔可夫模型，确定胎儿基因组中特定区域的碱基信息。由此，可以有效地对胎儿基因组的遗传信息进行产前检测。

[0022] 下面参考图 1，对利用惠特比算法借助隐马尔可夫模型进行分析的原理进行详细描述：

[0023] 在本文中所使用的术语“胎儿遗传相关个体”指的是在遗传意义上，与胎儿之间具有亲缘关系的个体，例如根据本发明的实施例，可以采用的“胎儿遗传相关个体”为胎儿的亲代例如父母。由此，子代的基因组形成过程，相当于亲代基因组的一次随机重组（即连锁互换单倍体型重组，以及配子的随机组合）。对于孕期血浆，假若我们将胎儿的单倍型（父母单倍型的重组型）作为隐含状态 (hidden states)，可以将血浆的测序数据当做观察序列 (observations)，借助先验数据推算出状态转移概率 (transition probabilities)、观察序列概率分布 (observation symbol probabilities) 和初始状态概率分布 (initial state distribution)，我们则可以通过惠特比算法 (Viterbi algorithm) 推断出最可能的胎儿单倍型组合，从而获得更多胎儿的信息。

[0024] 详细分析步骤如下：

[0025] 记号：

[0026] I. 需要检测的位点数为 N。

[0027] II. 父母的单倍型分别记为 $FH = \{fh_0, fh_1\}$ 和 $MH = \{mh_0, mh_1\}$，

[0028] 其中，

[0029] $mh_k = \{m_{1,k}, \ldots, m_{i,k}, \ldots, m_{N,k}\}$，$fh_k = \{f_{1,k}, \ldots, f_{i,k}, \ldots, f_{N,k}\}$，

[0030] $\forall fh_{i,k}, mh_{i,k} \in \{A, C, G, T\}$，

[0031] $k \in \{0,1\}$，$i = 1, 2, 3, \ldots, N$。

[0032] III. 将未知的胎儿单倍型记为 $H = \{h_0, h_1\}$，特别地，h_0 和 h_1 分别遗传自母亲和父亲。

[0033] $h_0 = \{m_{1,x_1}, \ldots, m_{i,x_i}, \ldots, m_{N,x_N}\}$，$h_1 = \{f_{1,y_1}, \ldots, f_{i,y_i}, \ldots, f_{N,y_N}\}$，

[0034] 其中，$x_i \in \{0,1\}$，$y_i \in \{0,1\}$，

[0035] 下标 x_i 和 y_i 组成的序列对，$q_i = \{x_i, y_i\}$ 组成了我们需要解码的隐藏状态，

[0036] 而所有可能出现的隐藏状态组成了集合 Q。

[0037] IV. 测序数据记为，$S = \{s_1, \ldots, s_i, \ldots, s_N\}$，

[0038] 其中，$s_i = \{n_{i,A}, n_{i,C}, n_{i,G}, n_{i,G}\}$，代表此位点的测序信息，包含了 ACGT 四种碱基的数量。

[0039] V. 平均胎儿浓度和平均测序错误率分别记为 ε 和 e。

[0040] 第一步，构建初始状态概率分布向量，以及单倍体重组转移矩阵：

[0041] I. 初始状态概率分布记为 $\pi = \{\pi_j\}$ $(j \in Q)$。

[0042] 根据本发明的实施例，在没有参考数据的情况下，可以设$\pi_j = \Pr(q_1 = j) \triangleq \frac{1}{4}$，即每种隐藏状态在第一个位点出现的可能性相等。

[0043] II. 根据本发明的实施例，记单倍体重组概率为 $p_r = re/N$，其中 re 代表人类配子基因组重组平均次数，为先验数据在 25 到 30 之间。

[0044] III. 根据本发明的实施例，记单倍体重组转移矩阵记为 $A = \{a_{jk}\}(j,k \in Q)$，其中 a_{jk} 为隐藏状态转移的概率，即

[0045] $$a_{jk} = \Pr(q_i = k | q_{i-1} = j) = \begin{cases} (1-p_r)^2 & x_i = x_{i-1}, y_i = y_{i-1} \\ (1-p_r)\cdot p_r & x_i = x_{i-1}, y_i \neq y_{i-1} \ \ or \ \ x_i \neq x_{i-1}, y_i = y_{i-1} \\ p_r^2 & x_i \neq x_{i-1}, y_i \neq y_{i-1} \end{cases},$$

[0046] 胎儿单倍型 $h_0 = \{m_{1,x_1}, ..., m_{i,x_i}, ..., m_{N,x_N}\}$，$h_1 = \{f_{1,y_1}, ..., f_{i,y_i}, ..., f_{N,y_N}\}$ 的下标 x_i 和 y_i 组成的序列对，$q_i = \{x_i, y_i\}$ 组成了我们需要解码的隐藏状态。举个例子，$x_i = 0$ 代表，“母源性染色体上，对应基因座上等位基因型为 $m_{i,0}$”。

[0047] 第二步，构建观察序列概率矩阵：

[0048] 根据本发明的实施例，记观察序列概率矩阵为 $B = \{b_{i,j}(s_i)\}(i = 1,2,3,...,N, j \in Q)$，其中 $b_{i,j}(s_i)$ 代表“在位点 i，考虑母亲单倍型和胎儿单倍型（状态 $j, j = \{x_i, y_i\}$）时，观测到这种测序信息的可能性”，即

[0049] $$\begin{aligned} b_{i,j}(s_i) &= \Pr(s_i | q_i = j, \{m_0, m_1\}) \\ &= \frac{(n_{i,A}+n_{i,C}+n_{i,G}+n_{i,T})!}{n_{i,A}!n_{i,C}!n_{i,G}!n_{i,T}!} \cdot (P_{i,A})^{n_{i,A}} \cdot (P_{i,C})^{n_{i,C}} \cdot (P_{i,G})^{n_{i,G}} \cdot (P_{i,T})^{n_{i,T}} \end{aligned},$$

[0050] 其中 $P_{i,base}$ 代表“在位点 i，考虑母亲单倍型和胎儿单倍型（状态 $j, j = \{x_i, y_i\}$）时，该碱基出现的可能性”，即

[0051] $$\begin{aligned} P_{i,base} &= \Pr(base | q_i = j, \{m_0, m_1\}) \\ &= \sum_{k \in \{0,1\}} \frac{1}{2}(1-\varepsilon)\Delta(base, m_k) + \frac{1}{2}\varepsilon \cdot \Delta(base, m_{x_i}) + \frac{1}{2}\varepsilon \cdot \Delta(base, f_{y_i}) \end{aligned},$$

[0052] 其中指示函数

[0053] $$\Delta(x,y) = \begin{cases} 1-e & x = y \\ e/3 & x \neq y \end{cases}。$$

[0054] 此步是进行 HMM 参数，每个位点的观察序列概率分布 $b_{i,j}(s_i)$ 计算，即计算每个位点上不同胎儿单倍型（隐藏状态）下，血浆出现当前测序数据（观察序列）的可能性。

[0055] 第三步，构建局部概率矩阵，和逆向指针（下面以一维局部概率矩阵构建为例）：

[0056] 定义局部概率 $\delta_i(q_i) = \left(\max_{q_{i-1} \in Q} \delta_i(q_i) \cdot a_{q_{i-1}q_i}\right) \cdot b_{i,q_i}(s_i)$

Pg. 704

[0057] 定义逆向指针 $\Psi_i(q_i)=\arg\max_{q_{i-1}\in Q}\delta_i(q_i)\cdot a_{q_{i-1}q_i}$

[0058] 这里所使用的术语局部概率 $\delta_i(q_i)$ 和逆向指针 $\Psi_i(q_i)$ 都是沿用Viterbi算法的经典定义。关于该参数的定义的详细描述，可以参见LawrenceR.Rabiner,PROCEEDINGS OF THE IEEE,Vol.77,No.2,1989年2月，通过参照将其全文并入本文。

[0059] 第四步，确定最终状态，并回溯最优路径：

[0060] 确定最终状态，$q_N^*=\arg\max_{q_N\in Q}\delta_N(q_N)$

[0061] 按照逆向指针回溯最优路径，即最可能胎儿单基因型 $q_i^*=\Psi_i(q_i)\ (i=1,2,3,...,N-1)$。

[0062] 第五步，输出结果。

[0063] 由此，能够有效地对胎儿基因组的序列进行分析。相比其他已有的产前检测技术方法，本方法有以下技术优势，主要体现在准确性和可获得的遗传信息量上：

[0064] 1)根据本发明实施例，检测的位点不仅针对父源性位点，对母源性位点，即母亲杂合位点，同样可以很好地检测出来胎儿是否遗传了母亲的致病位点，检测准确率可高达95%以上，且可以检测多种变异类型，扩大了疾病检测的范围。

[0065] 2)根据本发明实施例，不但可以通过一次测序获得多个位点、多种疾病的信息，对于一些在孕妇血浆中覆盖程度较低，单纯提高测序深度依然无法准确判定的基因序列，可以通过本方法推断得到，结果准确可靠。

[0066] 3)根据本发明实施例，可进行遗传疾病作图，对于一些连锁相关疾病，可通过其他位点的信息直接推断出来，一次可获得的信息量大，对临床检测更加具有指导意义。

[0067] 另外，根据本发明实施例，本发明的确定胎儿基因组中预定区域碱基信息的方法，不仅限于SNP或者STR等某一种遗传多态性位点，对所有的遗传多态性位点均可适用，且可以多种位点同时使用，以便互相验证。除了可进行产前无创检测胎儿基因组信息，达到疾病检测的目的，还可以进行无创产前亲子鉴定，在孩子出生前判定孩子父亲身份，为一些涉及抚养责任和义务、财产纠纷、性侵案等协助侦破。

[0068] 用于确定胎儿基因组中预定区域碱基信息的系统

[0069] 在本发明的又一方面，本发明提出了一种用于确定胎儿基因组中预定区域核酸序列的系统。根据本发明的实施例，参考图2，该系统1000可以包括：文库构建装置100、测序装置200以及分析装置400。

[0070] 根据本发明的实施例，文库构建装置100适于针对胎儿基因组DNA样本，构建测序文库。根据本发明的实施例，测序装置200与文库构建装置100相连，并且适于对所构建的测序文库进行测序，以便获得测序结果，所得到的测序结果由多个测序数据构成。根据本发明的实施例，还可以进一步包括DNA样本分离装置，该DNA样本分离装置适于从孕妇外周血中提取胎儿基因组DNA样本。由此，该系统可以适用于进行无创产前检测。

[0071] 根据本发明的实施例，任选地，还可以包括比对装置300。根据本发明的实施例，比对装置300与测序装置200相连，并且适于将所得到的测序结果与参照序列进行比对，以便确定与预定区域对应的测序数据。根据本发明的实施例，可以用于进行测序的方法和设备并不受特别限制，包括但不限于双脱氧链终止法；优选高通量的测序方法，由此，能够利用这些测序装置的高通量、深度测序的特点，进一步提高测序效率。从而，提高后续对测

Pg. 705

序数据进行分析，尤其是统计检验分析时的精确性和准确度。所述高通量的测序方法包括但不限于第二代测序技术或者是单分子测序技术。所述第二代测序平台 (MetzkerML. Sequencing technologies-the next generation. Nat Rev Genet. 2010Jan ;11(1):31-46) 包括但不限于 Illumina-Solexa(GA™, HiSeq2000™等)、ABI-Solid 和 Roche-454(焦磷酸测序)测序平台；单分子测序平台(技术)包括但不限于 Helicos 公司的真实单分子测序技术 (True Single Molecule DNA sequencing), Pacific Biosciences 公司单分子实时测序 (single molecule real-time(SMRT™)), 以及 Oxford Nanopore Technologies 公司的纳米孔测序技术等 (Rusk, Nicole(2009-04-01). Cheap Third-Generation Sequencing. Nature Methods6(4):244-245)。随着测序技术的不断进化，本领域技术人员能够理解的是还可以采用其他的测序方法和装置进行全基因组测序。根据本发明的具体示例，可以利用选自 Illumina-Solexa、ABI-SOLiD、Roche-454 和单分子测序装置的至少一种对所述全基因组测序文库进行测序。根据本发明的实施例，可以使用的参照序列的类型并不受特别限制，可以为任何含有感兴趣区域的已知序列。根据本发明的实施例，可以采用已知的人类参考基因组作为参照序列。例如，根据本发明的实施例，采用的人类参考基因组为 NCBI36.3, HG18。另外，根据本发明的实施例，进行比对的方法并不受特别限制。根据本发明的具体实施例，可以采用 SOAP 进行比对。

[0072] 根据本发明的实施例，分析装置 400 适于基于所述胎儿的测序结果，结合胎儿遗传相关个体的遗传信息，根据隐马尔可夫模型，确定所述预定区域的碱基信息。

[0073] 根据本发明的实施例，惠特比算法采用 0.25 作为初始状态概率分布，采用 re/N 作为重组概率，其中 re ＝ 25 ～ 30，优选 25，N 为所述预定区域的长度，

[0074] 采用

[0075] $$a_{jk}=\Pr(q_i=k|q_{i-1}=j)=\begin{cases}(1-p_r)^2 & x_i=x_{i-1},y_i=y_{i-1}\\(1-p_r)\cdot p_r & x_i=x_{i-1},y_i\neq y_{i-1}\ \ or\ \ x_i\neq x_{i-1},y_i=y_{i-1}\\p_r^2 & x_i\neq x_{i-1},y_i\neq y_{i-1}\end{cases}$$

[0076] 作为重组转移矩阵，p_r＝ re/N。

[0077] 根据本发明的实施例，将所述测序结果与参照序列进行比对，以便确定与所述预定区域对应的测序数据进一步包括按照下列公式确定概率最高的碱基：

[0078] $$P_{i,base}=\sum_{k\in\{0,1\}}\frac{1}{2}(1-\varepsilon)\Delta(base,m_k)+\frac{1}{2}\varepsilon\cdot\Delta(base,m_{x_i})+\frac{1}{2}\varepsilon\cdot\Delta(base,f_{y_i})$$

[0079] 其中，

[0080] $$\Delta(x,y)=\begin{cases}1-e & x=y\\e/3 & x\neq y\end{cases}$$

。

[0081] 关于数据分析部分，前面已经进行了详细描述，也当然地适用于确定胎儿基因组中预定区域核酸序列的系统。不再赘述。

[0082] 由此，利用该系统，能够有效地实施前面所述的确定胎儿基因组中预定区域核酸序列的方法，可以通过例如惠特比算法 (Viterbi algorithm)，借助隐马尔可夫模型，确定

胎儿基因组中特定区域的碱基信息，由此，可以有效地对胎儿基因组的遗传信息进行产前检测。

[0083] 此外，根据本发明的实施例，预定区域为已知存在遗传多态性的位点，而遗传多态性为选自单核苷酸多态性和STR的至少一种。

[0084] 在本文中所述的术语“相连”应作广义理解，既可以是直接相连，也可以是间接相连，只要能够实现上述功能上的衔接即可。

[0085] 需要说明的是，本领域技术人员能够理解，在前面所描述的确定胎儿基因组中预定区域核酸序列的方法的特征和优点也适合于确定胎儿基因组中预定区域核酸序列的系统，为描述方便，不再详述。

[0086] 计算机可读介质

[0087] 在本发明的又一方面，本发明提出了一种计算机可读介质。根据本发明的实施例，计算机可读介质上存储有指令，所述指令适于被处理器执行以便基于胎儿的测序结果，结合胎儿遗传相关个体的遗传信息，根据隐马尔可夫模型，确定所述预定区域的碱基信息。由此，利用该计算机可读介质，能够有效地实施前面所述的方法，从而可以通过例如惠特比算法(Viterbi algorithm)，借助隐马尔可夫模型，确定胎儿基因组中特定区域的碱基信息，由此，可以有效地对胎儿基因组的遗传信息进行产前检测。

[0088] 根据本发明的实施例，指令适于按照惠特比算法，根据隐马尔可夫模型，确定所述预定区域的碱基信息。根据本发明的实施例，在所述惠特比算法中，采用0.25作为初始状态概率分布，采用re/N作为重组概率，其中re＝25～30，优选25，N为所述预定区域的长度，

[0089] 采用

[0090] $$a_{jk}=\Pr(q_i=k|q_{i-1}=j)=\begin{cases}(1-p_r)^2 & x_i=x_{i-1},y_i=y_{i-1}\\ (1-p_r)\cdot p_r & x_i=x_{i-1},y_i\neq y_{i-1}\ \ or\ \ x_i\neq x_{i-1},y_i=y_{i-1}\\ p_r^2 & x_i\neq x_{i-1},y_i\neq y_{i-1}\end{cases}$$

[0091] 作为重组转移矩阵，其中，p_r＝rc/N。

[0092] 根据本发明的实施例，所述指令将所述测序结果与参照序列进行比对，以便确定与所述预定区域对应的测序数据进一步包括按照下列公式确定概率最高的碱基：

[0093] $$P_{i,base}=\sum_{k\in\{0,1\}}\frac{1}{2}(1-\varepsilon)\Delta(base,m_k)+\frac{1}{2}\varepsilon\cdot\Delta(base,m_{x_i})+\frac{1}{2}\varepsilon\cdot\Delta(base,f_{y_i})$$

[0094] 其中，

[0095] $$\Delta(x,y)=\begin{cases}1-e & x=y\\ e/3 & x\neq y\end{cases}$$

。

[0096] 关于数据分析部分，前面已经进行了详细描述，也当然地适用于确定胎儿基因组中预定区域核酸序列的系统。不再赘述。

[0097] 此外，根据本发明的实施例，预定区域为已知存在遗传多态性的位点，而遗传多态性为选自单核苷酸多态性和STR的至少一种。

[0098] 就本说明书而言，“计算机可读介质”可以是任何可以包含、存储、通信、传播或传输程序以供指令执行系统、装置或设备或结合这些指令执行系统、装置或设备而使用的装置。计算机可读介质的更具体的示例（非穷尽性列表）包括以下：具有一个或多个布线的电连接部（电子装置），便携式计算机盘盒（磁装置），随机存取存储器（RAM），只读存储器（ROM），可擦除可编辑只读存储器（EPROM 或闪速存储器），光纤装置，以及便携式光盘只读存储器（CDROM）。另外，计算机可读介质甚至可以是可在其上打印所述程序的纸或其他合适的介质，因为例如可以通过对纸或其他介质进行光学扫描，接着进行编辑、解译或必要时以其他合适方式进行处理来以电子方式获得所述程序，然后将其存储在计算机存储器中。

[0099] 应当理解，本发明的各部分可以用硬件、软件、固件或它们的组合来实现。在上述实施方式中，多个步骤或方法可以用存储在存储器中且由合适的指令执行系统执行的软件或固件来实现。例如，如果用硬件来实现，和在另一实施方式中一样，可用本领域公知的下列技术中的任一项或他们的组合来实现：具有用于对数据信号实现逻辑功能的逻辑门电路的离散逻辑电路，具有合适的组合逻辑门电路的专用集成电路，可编程门阵列（PGA），现场可编程门阵列（FPGA）等。

[0100] 本技术领域的普通技术人员可以理解实现上述实施例方法携带的全部或部分步骤是可以通过程序来指令相关的硬件完成，所述的程序可以存储于一种计算机可读存储介质中，该程序在执行时，包括方法实施例的步骤之一或其组合。

[0101] 此外，在本发明各个实施例中的各功能单元可以集成在一个处理模块中，也可以是各个单元单独物理存在，也可以两个或两个以上单元集成在一个模块中。上述集成的模块既可以采用硬件的形式实现，也可以采用软件功能模块的形式实现。所述集成的模块如果以软件功能模块的形式实现并作为独立的产品销售或使用时，也可以存储在一个计算机可读取存储介质中。

[0102] 下面将结合实施例对本发明的方案进行解释。本领域技术人员将会理解，下面的实施例仅用于说明本发明，而不应视为限定本发明的范围。实施例中未注明具体技术或条件的，按照本领域内的文献所描述的技术或条件（例如参考 J. 萨姆布鲁克等著，黄培堂等译的《分子克隆实验指南》，第三版，科学出版社）或者按照产品说明书进行。所用试剂或仪器未注明生产厂商者，均为可以通过市购获得的常规产品，例如可以采购自 Illumina 公司。

[0103] 一般方法

[0104] 本发明实施例的主要步骤包括：

[0105] 1）无创采取含有胎儿遗传物质的孕妇样品，提取其中含有的遗传物质。

[0106] 2）胎儿家庭成员如父母和外祖父母等基因组 DNA 提取和纯化。

[0107] 3）各遗传物质根据不同测序平台测序要求进行文库构建。

[0108] 4）测序获得的数据进行过滤，过滤条件根据质量值、接头污染等来设定。

[0109] 5）获得的高质量序列根据需要进行组装处理，组装结果与人类基因组参考序列进行比对。获得唯一比对的序列，带入模型进行分析。

[0110] 分析模型：

[0111] 记号：

[0112] I. 需要检测的位点数为 N。

[0113] II. 父母的单倍型分别记为 $FH = \{fh_0, fh_1\}$ 和 $MH = \{mh_0, mh_1\}$，

[0114] 其中，

[0115] $mh_k = \{m_{1,k}, \ldots, m_{i,k}, \ldots, m_{N,k}\}$，$fh_k = \{f_{1,k}, \ldots, f_{i,k}, \ldots, f_{N,k}\}$，

[0116] $\forall fh_{i,k}, mh_{i,k} \in \{A, C, G, T\}$，

[0117] $k \in \{0, 1\}$，$i = 1, 2, 3, \ldots, N$。

[0118] III. 将未知的胎儿单倍型记为 $H = \{h_0, h_1\}$，特别地，h_0 和 h_1 分别遗传自母亲和父亲。

[0119] $h_0 = \{m_{1,x_1}, \ldots, m_{i,x_i}, \ldots, m_{N,x_N}\}$，$h_1 = \{f_{1,y_1}, \ldots, f_{i,y_i}, \ldots, f_{N,y_N}\}$，

[0120] 其中，$x_i \in \{0, 1\}$，$y_i \in \{0, 1\}$，

[0121] 下标 x_i 和 y_i 组成的序列对，$q_i = \{x_i, y_i\}$ 组成了我们需要解码的隐藏状态，

[0122] 而所有可能出现的隐藏状态组成了集合 Q。

[0123] IV. 测序数据记为，$S = \{s_1, \ldots, s_i, \ldots, s_N\}$

[0124] 其中，$s_i = \{n_{i,A}, n_{i,C}, n_{i,G}, n_{i,G}\}$，代表此位点的测序信息，包含了 ACGT 四种碱基的数量。

[0125] V. 平均胎儿浓度和平均测序错误率分别记为 ε 和 c。

[0126] 第一步，构建初始状态概率分布向量，以及单倍体重组转移矩阵：

[0127] I. 初始状态概率分布记为 $\pi = \{\pi_j\}(j \in Q)$，

[0128] 根据本发明的实施例，在没有参考数据的情况下，可以设 $\pi_j = \Pr(q_1 = j) \triangleq \frac{1}{4}$，即每种隐藏状态在第一个位点出现的可能性相等。

[0129] II. 根据本发明的实施例，记单倍体重组概率为 $p_r = re/N$，其中 re 代表人类配子基因组重组平均次数，为先验数据在 25 到 30 之间。

[0130] III. 根据本发明的实施例，记单倍体重组转移矩阵记为 $A = \{a_{jk}\}(j, k \in Q)$，其中 a_{jk} 为隐藏状态转移的概率，即

[0131] $$a_{jk} = \Pr(q_i = k | q_{i-1} = j) = \begin{cases} (1-p_r)^2 & x_i = x_{i-1}, y_i = y_{i-1} \\ (1-p_r) \cdot p_r & x_i = x_{i-1}, y_i \neq y_{i-1} \quad or \quad x_i \neq x_{i-1}, y_i = y_{i-1} \\ p_r^2 & x_i \neq x_{i-1}, y_i \neq y_{i-1} \end{cases}$$

[0132] 第二步，构建观察序列概率矩阵：

[0133] 根据本发明的实施例，记观察序列概率矩阵为 $B = \{b_{i,j}(s_i)\}(i = 1, 2, 3, \ldots, N, j \in Q)$，其中 $b_{i,j}(s_i)$ 代表“在位点 i，考虑母亲单倍型和胎儿单倍型（状态 j）时，观测到这种测序信息的可能性”，即

[0134] $$b_{i,j}(s_i) = \Pr(s_i | q_i = j, \{m_0, m_1\}) = \frac{(n_{i,A} + n_{i,C} + n_{i,G} + n_{i,T})!}{n_{i,A}! n_{i,C}! n_{i,G}! n_{i,T}!} \cdot (P_{i,A})^{n_{i,A}} \cdot (P_{i,C})^{n_{i,C}} \cdot (P_{i,G})^{n_{i,G}} \cdot (P_{i,T})^{n_{i,T}}$$

[0135] 其中 $P_{i,base}$ 代表“在位点 i，考虑母亲单倍型和胎儿单倍型（状态 j）时，该碱基出

现的可能性”，即

[0136] $$P_{i,base} = \Pr\left(base \middle| q_i = j, \{m_0, m_1\}\right) = \sum_{k\in\{0,1\}} \frac{1}{2}(1-\varepsilon)\Delta\left(base, m_k\right) + \frac{1}{2}\varepsilon\cdot\Delta\left(base, m_{x_i}\right) + \frac{1}{2}\varepsilon\cdot\Delta\left(base, f_{y_i}\right),$$

[0137] 其中指示函数

[0138] $$\Delta(x,y) = \begin{cases} 1-e & x = y \\ e/3 & x \neq y \end{cases}。$$

[0139] 第三步，构建局部概率矩阵，和逆向指针（下面以一维局部概率矩阵构建为例）：

[0140] 定义局部概率 $\delta_i(q_i) = \left(\max_{q_{i-1}\in Q}\delta_i(q_i)\cdot a_{q_{i-1}q_i}\right)\cdot b_{i,q_i}(s_i)$

[0141] 定义逆向指针 $\Psi_i(q_i) = \arg\max_{q_{i-1}\in Q}\delta_i(q_i)\cdot a_{q_{i-1}q_i}$

[0142] 第四步，确定最终状态状态，并回溯最优路径：

[0143] 确定最终状态，$q_N^* = \arg\max_{q_N\in Q}\delta_N(q_N)$，

[0144] 根据逆向指针回溯最优路径，即最可能胎儿单基因型 $q_i^* = \Psi_i(q_i)$ $(i = 1,2,3,\ldots,N-1)$。

[0145] 第五步，输出结果。

[0146] 实施例 1

[0147] 样品收集及处理：

[0148] (1) 所取样品包括一个家庭中父亲和母亲孕期的外周血，胎儿出生后取脐带血，以EDTA 抗凝管收集，Oragene® DNA 唾液采集 /DNA 纯化试剂盒 OG-250 采集祖父母和外祖父母唾液。

[0149] (2) 祖父母和外祖父母唾液 DNA 提取后用 Infinium® HD Human610-Quad BeadChip 基因芯片进行分型。

[0150] (3) 取母亲孕期外周血，1600g，4℃离心 10 分钟，将血细胞和血浆分开，血浆再以 16000g，4℃离心 10 分钟，进一步去除残留的白细胞。孕妇血浆用 TIANamp Micro DNA Kit(TIANGEN) 提取 DNA，得到母亲和胎儿基因组 DNA 混合物，并且从全血分离血浆后剩余的白细胞中提取母亲基因组 DNA。将所得到的血浆 DNA 根据 Illumia® 公司 HiSeq2000™ 测序仪的上机要求进行建库，构建好的文库经 Agilent® Bioanalyzer2100 检测片段分布范围符合要求，再经过 Q-PCR 方法对两个文库进行定量，合格后 Illumina® HiSeq2000™ 测序仪测序，测序循环数为 PE101index（即双向 101bp index 测序），其中仪器的参数设置及操作方法都按照 Illumina® 操作手册（可由 http://www.illumina.com/support/documentation.ilmn 获取）。

[0151] (4) 父亲外周血、母亲外周血白细胞和胎儿脐带血则直接用 TIANamp Micro DNA Kit(TIANGEN) 提取试剂盒提取基因组 DNA。

[0152] 除血浆DNA样品外，将所获得的所有DNA样品，需用Covaris™打断仪打断至500bp大小的片段。将获得的DNA片段以及血浆DNA样品根据Illumia®公司HiSeq2000™测序仪的上机要求进行建库，具体步骤如下：

[0153] 末端修复反应体系：

[0154]

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

[0155] 20℃反应30分钟后，使用PCR Purification Kit(QIAGEN)回收末端修复产物。将所得到的产物最后溶于34μl的EB缓冲液中。

[0156] 末端添加碱基A反应体系：

[0157]

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

[0158] 37℃温育30分钟后，经MinElute® PCR Purification Kit(QIAGEN)纯化并溶于12μl的EB中。

[0159] 接头连接反应体系：

[0160]

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

[0161] 20℃反应15分钟后，使用PCR Purification Kit(QIAGEN)回收连接产物。将所得到的产物最后溶于32μl的EB缓冲液中。

[0162]

PCR 反应体系：	
接头连接反应产物	10 μl
Phusion DNA Polymerase Mix	25 μl
PCR 引物 (10 pmol/μl)	1 μl
Index N(10 pmol/μl)	1 μl
超纯水	13 μl

[0163] 反应程序如下：

[0164]

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

[0165] 使用 PCR Purification Kit(QIAGEN) 回收 PCR 产物。样品最后溶于 50μl 的 EB 缓冲液中。

[0166] 构建好的文库经 Agilent® Bioanalyzer2100 检测片段分布范围符合要求，再经过 Q-PCR 方法对两个文库进行定量，合格后，用 Illumina® HiSeq2000™测序仪测序，测序循环数为 PE101index（即双向 101bp index 测序），其中仪器的参数设置及操作方法都按照 Illumina® 操作手册（可由 http://www.illumina.com/support/documentation.ilmn 获取）。

[0167] (5) 父母基因组测序分型：

[0168] a. 使用 SOAP2 将测序数据比对到人类参考基因组（版本为 NCBI36.3，HG18）。

[0169] b. 使用 SOAPsnp（南方汉族 (CHS) 家系数据使用的是千人计划数据）对数据进行一致序列 (consensus sequence，CNS) 构建。

[0170] c. 提取出标记位点的基因型。

[0171] (6) 父母单倍体型推断：

[0172] a. 构建含祖辈与父母基因型的群体基因型矩阵，即提取父母、祖辈和南方汉族家系在标记位点的基因型。

[0173] b. 使用 BEAGLE 对父母的单倍型进行推断。

[0174] (7) 胎儿单倍体型推断：

[0175] a. 用 SOAP2 将血浆测序数据比对到人类参考基因组（版本为 NCBI36.3，HG18）。

[0176] b. 构建初始状态概率向量，以及单倍体重组转移矩阵。

[0177] 构建初始状态概率向量：采取无参考数据模式，即各个初始状态概率相等，均为 0.25。

[0178] 单倍体重组转移矩阵：保守地，我们取 re = 25（其余按一般方法所述）。

[0179] c. 统计每个位点的测序信息，并构建观察序列概率矩阵（其余按一般方法所述）。

Pg. 712

[0180] d. 构建局部概率矩阵，和逆向指针（其余按一般方法所述）。

[0181] e. 确定最终状态状态，并回溯最优路径。

[0182] f. 输出。

[0183] 根据胎儿出生后的脐带血基因分型结果，我们的分类准确性统计如下：

[0184]

			母亲						合计		
			纯合			杂合					
			位点数	准确数	准确率	位点数	准确数	准确率	位点数	准确数	准确率
常染色体	父亲	纯合	199,552	199,552	100.00%	66,238	63,958	96.57%	265,790	263,520	99.15%
		杂合	65,409	64,735	98.97%	41,849	39,944	95.45%	107,258	104,679	97.60%
	合计		264,961	264,287	99.75%	108,087	103,912	96.14%	373,048	368,199	98.70%
X染色体			4,881	4,881	100.00%	1,718	1,478	86.03%	6,599	6,359	96.36%

[0185] 工业实用性

[0186] 本发明的确定胎儿基因组中预定区域碱基信息的方法、用于确定胎儿基因组中预定区域碱基信息的系统以及计算机可读介质，能够有效地应用于对胎儿基因组中预定区域的核酸序列进行分析。

[0187] 尽管本发明的具体实施方式已经得到详细的描述，本领域技术人员将会理解。根据已经公开的所有教导，可以对那些细节进行各种修改和替换，这些改变均在本发明的保护范围之内。本发明的全部范围由所附权利要求及其任何等同物给出。

[0188] 在本说明书的描述中，参考术语“一个实施例”、“一些实施例”、“示意性实施例”、“示例”、“具体示例”、或“一些示例”等的描述意指结合该实施例或示例描述的具体特征、结构、材料或者特点包含于本发明的至少一个实施例或示例中。在本说明书中，对上述术语的示意性表述不一定指的是相同的实施例或示例。而且，描述的具体特征、结构、材料或者特点可以在任何的一个或多个实施例或示例中以合适的方式结合。

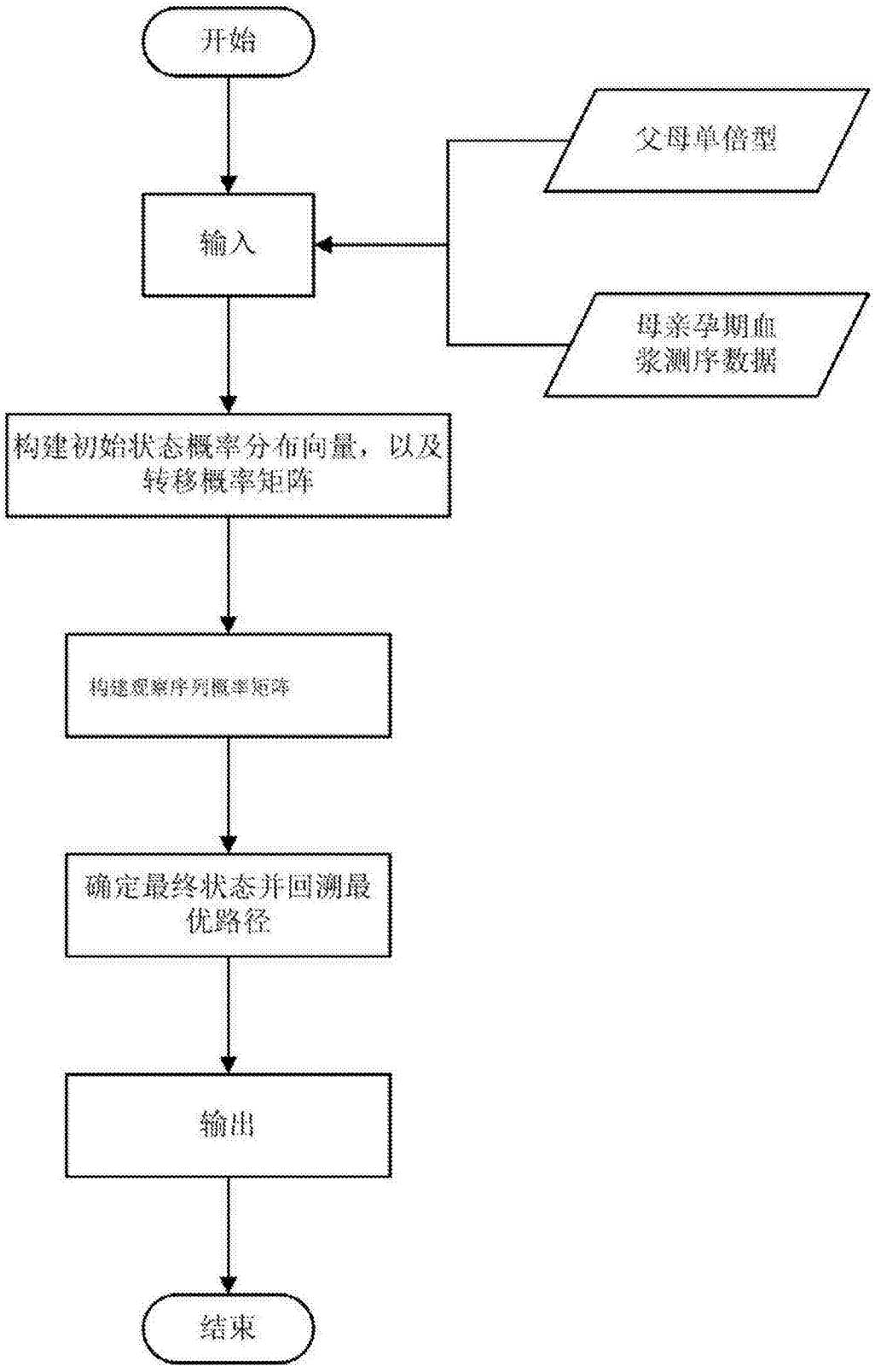

开始
父母单倍型
输入
母亲孕期血浆测序数据
构建初始状态概率分布向量，以及转移概率矩阵
构建观察序列概率矩阵
确定最终状态并回溯最优路径
输出
结束

图 1

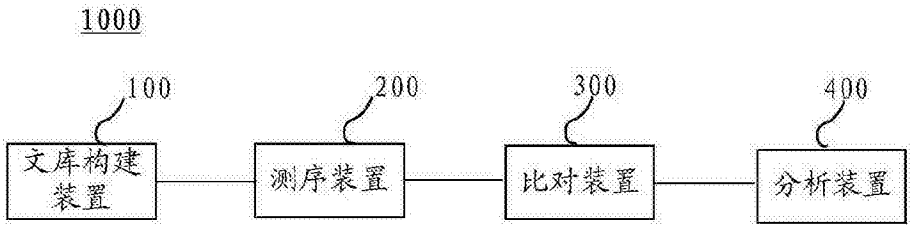


图 2

Espacenet

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

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

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

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

Classification: **- international:** *C12Q1/68*

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

C-sets

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

less

Application number: WO2012CN75478 20120514 Global Dossier

Priority number(s): WO2012CN75478 20120514

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

Abstract of WO2013170429 (A1)

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

(12) 按照专利合作条约所公布的国际申请

(19) 世界知识产权组织
国 际 局

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

WIPO | PCT

(10) 国际公布号
WO 2013/170429 A1

(51) 国际专利分类号:
C12Q 1/68 (2006.01)

(21) 国际申请号: PCT/CN2012/075478

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

(25) 申请语言: 中文

(26) 公布语言: 中文

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

(72) 发明人; 及

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

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

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

(81) 指定国 (除另有指明，要求每一种可提供的国家保护): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IS, JP, KE, KG, KM, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW。

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

本国际公布:

— 包括国际检索报告(条约第 21 条(3))。

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

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

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

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

WO 2013/170429 A1

确定胎儿基因组中预定区域碱基信息的方法、系统和计算机可读介质

优先权信息

无

技术领域

本发明涉及确定胎儿基因组中预定区域碱基信息的方法、系统和计算机可读介质。

背景技术

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

然而，目前的相关检测手段仍有待改进。

发明内容

本发明旨在至少解决现有技术中存在的技术问题之一。

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

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

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

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

附图说明

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

图 1 为根据本发明一个实施例的利用隐马尔可夫模型进行分析的流程示意图；以及

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

发明详细描述

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

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

确定胎儿基因组中预定区域碱基信息的方法

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

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

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

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

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

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

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

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

详细分析步骤如下：

记号：

I.需要检测的位点数为 N。

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

其中，

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

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

$k\in\{0,1\}$， $i=1,2,3,...,N$。

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

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

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

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

而所有可能出现的隐藏状态组成了集合 Q。

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

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

V. 平均胎儿浓度和平均测序错误率分别记为 ε 和 e。

第一步，构建初始状态概率分布向量，以及单倍体重组转移矩阵:

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

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

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

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

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

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

第二步，构建观察序列概率矩阵:

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

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

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

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

其中指示函数

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

。

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

第三步，构建局部概率矩阵，和逆向指针（下面以一维局部概率矩阵构建为例）:

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

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

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

第四步，确定最终状态，并回溯最优路径:

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

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

第五步，输出结果。

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

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

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

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

点的信息直接推断出来，一次可获得的信息量大，对临床检测更加具有指导意义。

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

用于确定胎儿基因组中预定区域碱基信息的系统

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

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

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

对。

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

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

采用

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

作为重组转移矩阵， p_r=re/N。

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

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

其中，

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

。

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

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

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

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

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

计算机可读介质

在本发明的又一方面，本发明提出了一种计算机可读介质。根据本发明的实施例，计

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

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

采用

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

作为重组转移矩阵，其中，p_r=re/N。

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

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

其中，

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

。

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

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

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

存储器中。

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

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

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

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

一般方法

本发明实施例的主要步骤包括:

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

分析模型:

记号:

I.需要检测的位点数为 N。

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

其中，

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

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

$k \in \{0,1\}$， $i = 1,2,3,...,N$。

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

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

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

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

而所有可能出现的隐藏状态组成了集合 Q。

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

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

V. 平均胎儿浓度和平均测序错误率分别记为 ε 和 e。

第一步，构建初始状态概率分布向量，以及单倍体重组转移矩阵：

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

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

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

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

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

第二步，构建观察序列概率矩阵：

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

信息的可能性”，即

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

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

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

其中指示函数

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

第三步，构建局部概率矩阵，和逆向指针（下面以一维局部概率矩阵构建为例）:

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

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

第四步，确定最终状态状态，并回溯最优路径:

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

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

第五步，输出结果。

实施例 1

样品收集及处理:

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

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

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

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

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

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

末端修复反应体系：

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

ddH_2O 补齐至 100 μl

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

末端添加碱基 A 反应体系：

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

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

接头连接反应体系：

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

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

PCR 反应体系：

接头连接反应产物	10 μl

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

反应程序如下:

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

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

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

（5）父母基因组测序分型:

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

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

c. 提取出标记位点的基因型。

（6）父母单倍体型推断:

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

b. 使用 BEAGLE 对父母的单倍型进行推断。

（7）胎儿单倍体型推断:

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

b. 构建初始状态概率向量，以及单倍体重组转移矩阵。

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

单倍体重组转移矩阵：保守地，我们取 $re = 25$（其余按一般方法所述）。

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

d. 构建局部概率矩阵，和逆向指针（其余按一般方法所述）。

e. 确定最终状态状态，并回溯最优路径。

f. 输出。

根据胎儿出生后的脐带血基因分型结果，我们的分类准确性统计如下:

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

工业实用性

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

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

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

权利要求书

1、一种确定胎儿基因组中预定区域碱基信息的方法，其特征在于，包括下列步骤：

针对胎儿基因组 DNA 样本，构建测序文库；

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

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

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

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

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

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

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

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

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

采用

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

作为重组转移矩阵，其中，p_r=re/N。

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

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

其中，

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

。

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

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

12、一种用于确定胎儿基因组中预定区域碱基信息的系统，其特征在于，包括：

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

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

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

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

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

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

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

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

采用

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

作为重组转移矩阵，其中，p_r=re/N。

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

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

其中，

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

。

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

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

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

采用

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

作为重组转移矩阵，其中，p_r=re/N。

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

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

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

其中，

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

。

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

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

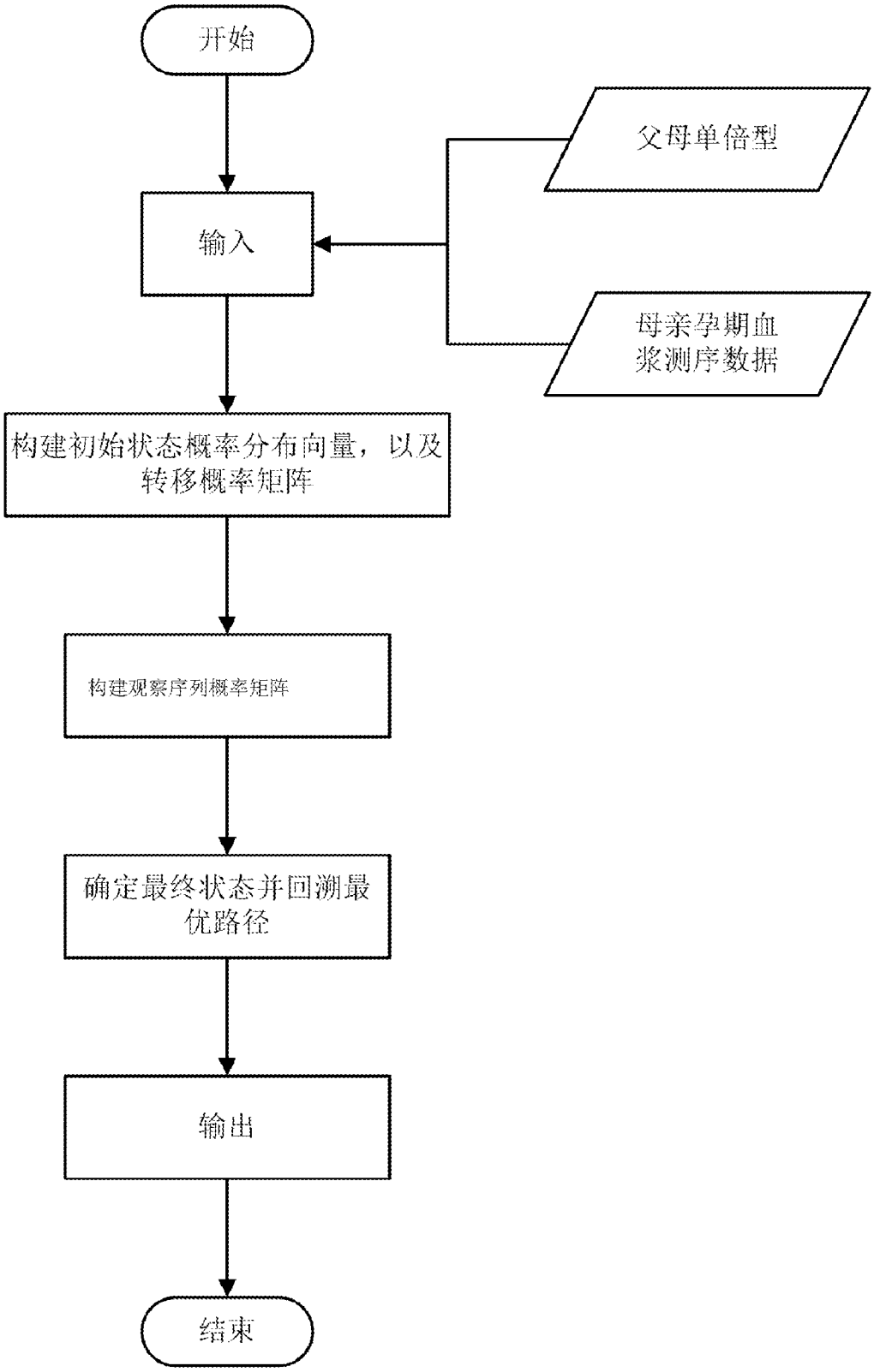

开始
父母单倍型
输入
母亲孕期血浆测序数据
构建初始状态概率分布向量，以及转移概率矩阵
构建观察序列概率矩阵
确定最终状态并回溯最优路径
输出
结束

图 1

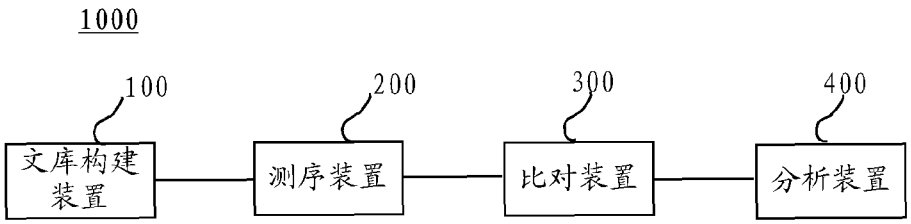


图 2

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2012/075478

A. CLASSIFICATION OF SUBJECT MATTER

C12Q 1/68 (2006.01) i

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

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

IPC: C12Q

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

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

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

C. DOCUMENTS CONSIDERED TO BE RELEVANT

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

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

* Special categories of cited documents:

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

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

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

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

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

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

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

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

"&"document member of the same patent family

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

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

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

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2012/075478

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

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

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

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

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

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

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

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

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

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

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

Remark on protest

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

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

☐ No protest accompanied the payment of additional search fees.

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2012/075478

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

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

国际检索报告

国际申请号
PCT/CN2012/075478

A. 主题的分类

C12Q 1/68 (2006.01) i

按照国际专利分类(IPC)或者同时按照国家分类和 IPC 两种分类

B. 检索领域

检索的最低限度文献(标明分类系统和分类号)

IPC: C12Q

包含在检索领域中的除最低限度文献以外的检索文献

在国际检索时查阅的电子数据库(数据库的名称，和使用的检索词（如使用）)

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

C. 相关文件

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

☐ 其余文件在 C 栏的续页中列出。　　☒ 见同族专利附件。

* 引用文件的具体类型:

"A" 认为不特别相关的表示了现有技术一般状态的文件

"E" 在国际申请日的当天或之后公布的在先申请或专利

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

"O" 涉及口头公开、使用、展览或其他方式公开的文件

"P" 公布日先于国际申请日但迟于所要求的优先权日的文件

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

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

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

"&" 同族专利的文件

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

国际检索报告

国际申请号
PCT/CN2012/075478

第II栏　某些权利要求被认为是不能检索的意见(续第 1 页第 2 项)

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

1. ☒ 权利要求：19-25

因为它们涉及不要求本单位进行检索的主题，即：

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

2. ☐ 权利要求：

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

具体地说：

3. ☐ 权利要求：

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

第III栏　缺乏发明单一性的意见(续第 1 页第 3 项)

本国际检索单位在该国际申请中发现多项发明，即：

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

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

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

具体地说，是权利要求：

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

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

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

☐ 缴纳附加检索费时未提交异议书。

国际检索报告
关于同族专利的信息

国际申请号
PCT/CN2012/075478

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

(12) 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.

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

2012380221 01 Sep 2016

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;

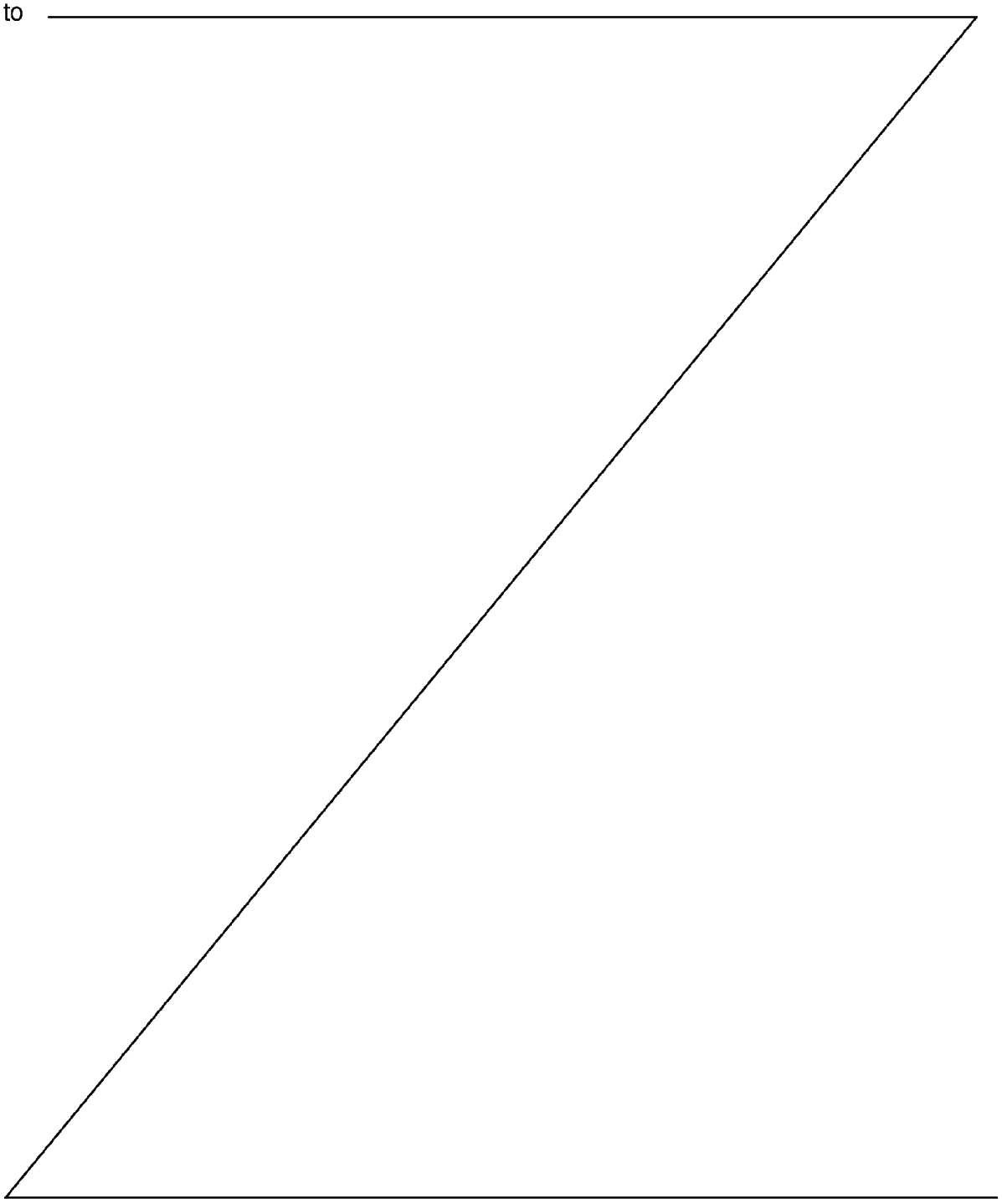
constructing a partial probability matrix, and a reversal cursor;

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

embodiments of the disclosure, the method may comprise: 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. A formation of offspring genome equals to a random recombination with parental generation's genome (i.e., an interchange of haplotype recombination, and a random combination of gametes). For pregnant plasma, if a fetal haplotype (a recombination of parental haplotypes) is assumed as hidden states, sequencing data of the plasma may be used as observations (observing sequence), transition probabilities, observation symbol probabilities and initial state distribution may be deduced in virtue of prior data, then the most possible fetal haplotype recombination may be determined using a hidden Markov Model based on Viterbi algorithm, so as to obtain more information of fetus prior to birth. Thus, according to embodiments of the present disclosure, in virtue of the hidden Markov Model, for example using the Viterbi algorithm, and referring to genetic information of a related individual, nucleic acid sequence of a predetermined region in a fetal genome may be determined, by which a prenatal genetic detection may be effectively performed with genetic information of fetal genome.

Embodiments of a second broad aspect of the present disclosure provide a system for determining base information of a predetermined region in a fetal genome. According to embodiments of the present disclosure, the system may comprise: 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. Using the system may effectively implement the above method of determining base information of a predetermined region in a fetal genome, which may determine nucleic acid sequence of a predetermined region in a fetal genome may be

determined in virtue of the hidden Markov Model, for example using the Viterbi algorithm, and referring to genetic information of a related individual, by which a prenatal genetic detection may be effectively performed with genetic information of the fetal genome.

Embodiments of a third broad aspect of the present disclosure provide a computer readable medium. According to embodiments of the present disclosure, the computer readable medium including a plurality of instructions is 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. Using the computer readable medium of the present disclosure may effectively execute the plurality of instructions by a processor, to determine a nucleic acid sequence of the predetermined region in the fetal genome in virtue of the hidden Markov Model, for example using the Viterbi algorithm based on the sequencing data of the fetus combining with genetic information of a related individual, by which prenatal genetic detection may be effectively performed with genetic information of the fetal genome.

Additional aspects and advantages of embodiments of present disclosure will be given in part in the following descriptions, become apparent in part from the following descriptions, or be learned from the practice of the embodiments of the present disclosure.

BRIEF DESCRIPTION OF THE DRAWINGS

These and other aspects and advantages of embodiments of the present disclosure will become apparent and more readily appreciated from the following descriptions made with reference the accompanying drawings, in which:

Fig.1 is a flow chart showing an analyzing process using a hidden Markov Model according to an embodiment of the present disclosure; and

Fig.2 is a schematic diagram showing a system for determining base information of a predetermined region in a fetal genome according to an embodiment of the present disclosure.

DETAILED DESCRIPTION

Reference will be made in detail to embodiments of the present disclosure. The same or similar elements and the elements having same or similar functions are denoted

by like reference numerals throughout the descriptions. The embodiments described herein with reference to drawings are explanatory, illustrative, and used to generally understand the present disclosure. The embodiments shall not be construed to limit the present disclosure.

It should note that terms such as "first" and "second" are used herein for purposes of description and are not intended to indicate or imply relative importance or significance. Thus, features defined with “first”, “second” may explicitly or implicitly include one or more the features. Furthermore, in the description of the present disclosure, unless otherwise stated, “a/the plurality of” means two or more.

Method of determining base information of a predetermined region in a fetal genome

In a first aspect of the present disclosure, there is provided a method of determining base information of a predetermined region in a fetal genome. According to embodiments of the present disclosure, the method may comprise:

firstly, constructing a sequencing library based on a genomic DNA sample of a fetus. According to embodiments of the present disclosure, source of the genomic DNA sample of the fetus is not subjected to any special restrictions. According to some embodiments of the present disclosure, any pregnant samples containing a nucleic acid of a fetal may be used. For example, according to embodiments of the present disclosure, the pregnant sample may be breast milk, urine and peripheral blood from a pregnant woman. In which, the pregnant peripheral blood is preferred. Using the pregnant peripheral blood as the source of the genomic DNA sample of the fetus may effectively realize obtaining the genomic DNA sample of the fetus by noninvasive sampling, by which the fetal genome may be effectively monitored in the premise of having no influence on normal development of fetal growth. As for methods and processes of constructing a sequencing library for the nucleic acid sample, a person skilled in the art may appropriately select depending on different sequencing technology. Detailed process may refer to procedure provided by sequencer manufacturer, such as Illumina Company, for example, refer to Multiplexing Sample Preparation Guide (Part#1005361; Feb 2010) or Paired-End SamplePrep Guide (Part#1005063; Feb 2010) from Illumina Company, which are incorporated herein for reference. According to embodiments of the present disclosure, methods and devices for extracting a nucleic acid from a biological sample are not

subjected to any special restrictions, which may be performed using a commercial nucleic acid extracting kit.

After being constructed, obtained sequencing library is applied to a sequencer, to obtain a corresponding sequencing result consisting of a plurality of sequencing data. According to embodiments of the present disclosure, methods and devices for sequencing are not subjected to any special restrictions, including but not limited to Chain Termination Method (Sanger); a high-throughput sequencing method is preferred. Thus, using characteristics being high-throughput and deep sequencing of these apparatus, efficiency may be further improved, by which precise and accuracy of subsequent analysis with sequencing data, such as statistical test, may be further improved. The high-throughput sequencing method includes but not limited to a Next-Generation sequencing technology or a single sequencing technology. The Next-Generation sequencing platform (Metzker ML. Sequencing technologies-the next generation. *Nat Rev Genet*. 2010 Jan; 11(1):31-46) includes but not limited to Illumina-Solexa (GATM, HiSeq2000TM, etc), ABI-Solid and Roche-454 (pyrosequencing) sequencing platform; the single sequencing platform (technology) includes but not limited to True Single Molecule DNA sequencing of Helicos Company, single molecule real-time (SMRT^{TM}) of Pacific Biosciences Company, and nonapore sequencing technology of Oxford Nanopore Technologies (Rusk, Nicole (2009-04-01). Cheap Third-Generation Sequencing. *Nature Methods* 6 (4): 244-245), etc. With gradual development of sequencing technology, a person skilled in the art may understand other sequencing methods and apparatuses may also be used for whole genome sequencing. According to specific examples of the present disclosure, the whole genome sequencing library may be subjected to sequencing by at least one selected from Illumina-Solexa, ABI-SOLiD, Roche-454 and a single molecule sequencing apparatus.

Optionally, after being obtained, the sequencing result may be aligned to a reference sequence, to determine sequencing data corresponding to the predetermined region. Term of "predetermined region" used herein should be broadly understood, referring to any region of a nucleic acid molecule containing a possible predetermined event. For SNP analysis, it may be a region containing SNP site. For analyzing chromosome aneuploidy, the predetermined region refers to entire or part of the chromosome to be analyzed, i.e., selecting sequencing data deriving from the chromosome. Methods of

selecting sequencing data deriving from a corresponding region in the sequencing result are not subjected to any special restrictions. According to embodiments of the present disclosure, all obtained sequencing data may be aligned to a reference sequence with a known nucleic acid, to obtain the sequencing data deriving from the predetermined region. In addition, according to embodiments of the present disclosure, the predetermined region may also be a plurality of dispersal points which are not discontinuous in a genome. According to embodiments of the present disclosure, a type of used reference sequence may be not subjected to any special restrictions, which may be any known sequences contained a target region. According to embodiments of the present disclosure, the reference sequence may use a known human reference genome. For example, according to embodiments of the present disclosure, the human reference genome is NCBI 36.3, HG18. In addition, according to embodiments of the present disclosure, alignment methods are not subjected to any special restrictions. According to specific examples, SOAP may be used for alignment.

Then, determining a part of a nucleic acid sequence of the predetermined region based on sequencing data corresponding to the predetermined region; and determining other parts of the nucleic acid sequence based on determined part of the nucleic acid sequence of the predetermined region using Viterbi algorithm, to obtain the nucleic acid sequence of the predetermined region. According to embodiments of the present disclosure, the base information of the predetermined region is determined based on the sequencing result of the fetus combining with genetic information of a related individual using a hidden Markov Model. According to embodiments of the present disclosure, the base information of the predetermined region is determined using the hidden Markov Model is performed based on Viterbi algorithm. Thus, a prenatal genetic detection may be effectively performed with genetic information of fetal genome.

Referring to Fig.1, a principal for analysis using Viterbi algorithm in virtue of a hidden Markov Model is descripted in details below:

In the genetic sense, term of “a related individual” refers to individuals having a genetic relationship with a fetus. For example, according to embodiments of the present disclosure, “a related individual” may be a rental generation of a fetus, such as parents. Thus, a formation of offspring genome equals to a random recombination with parental generation’s genome (i.e., an interchange of haplotype recombination, and a random

combination of gametes). For pregnant plasma, if a fetal haplotype (a recombination of parental haplotypes) is assumed as hidden states, sequencing data of the plasma may be used as observations (observing sequence), transition probabilities, observation symbol probabilities and initial state distribution may be deduced in virtue of prior data, then the most possible fetal haplotype recombination may be determined using a hidden Markov Model based on Viterbi algorithm, so as to obtain more information of fetus prior to birth.

Steps of analyzing are show below in details:

Marker:

I. the number of sites to be detected is N.

II. haplotypes of parents are respectively recorded as $FH = \{fh_0, fh_1\}$ and $MH = \{mh_0, mh_1\}$,

in which

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

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

$k \in \{0,1\}$, $i = 1,2,3,...,N$.

III. Unknown fetal haplotype is recorded as $H = \{h_0, h_1\}$, particularly, h_0 and h_1 respectively represent inheriting from mother and father.

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

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

Subscripts x_i and y_i respectively present sequence pairs, and $q_i = \{x_i, y_i\}$ represents the hidden states which need to be decoded.

While, all hidden states possible presenting constitutes a set Q.

IV. Sequencing data is recorded as $S = \{s_1,...,s_i,...,s_N\}$

in which $s_i = \{n_{i,A}, n_{i,C}, n_{i,G}, n_{i,G}\}$ represents sequencing information of a site, containing the number of four bases, A, C, T and G.

V. A mean fetal concentration and a mean sequencing error rate are respectively

recorded as ε and e.

Step 1, constructing a probability distribution vector of an initial state and a transition matrix of haplotypes recombination:

I. The probability distribution of the initial states is recorded as $\pi = \{\pi_j\}$ $(j \in Q)$.

According to embodiments of the present disclosure, under a circumstance of having no reference data, it may assume that $\pi_j = \Pr(q_1 = j) \triangleq \frac{1}{4}$, i.e., possibilities of each hidden state presenting at the first site are equal.

II. According to embodiments of the present disclosure, a probability of haplotype recombination is recorded as $p_r = re/N$, in which re represents a mean times of human gamete recombinations, with a prior data ranging from 25 to 30.

III. According to embodiments of the present disclosure, a transition matrix of haplotypes recombination is recorded as $A = \{a_{jk}\}$ $(j,k \in Q)$, in which a_{jk} represents a probability of hidden states transition, i.e.,

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

Subscripts x_i and y_i of fetal haplotypes $h_0 = \{m_{1,x_1},...,m_{i,x_i},...,m_{N,x_N}\}$ and $h_1 = \{f_{1,y_1},...,f_{i,y_i},...,f_{N,y_N}\}$ constitute a sequence pair, $q_i = \{x_i, y_i\}$ constitute the hidden states to be encoded. For example, $x_i = 0$ represents "in a maternal chromosome, an allele in the corresponding locus is $m_{i,0}$".

Step 2, constructing a probability matrix of observations:

According to embodiments of the present disclosure, the probability matrix of observations is recorded as $B = \{b_{i,j}(s_i)\}$ $(i = 1,2,3,...,N,\ j \in Q)$, in which $b_{i,j}(s_i)$ represents "an observed possibility of this sequencing information in a site i, considering maternal haplotype and fetal haplotype (state j, j={x_i, y_i})", i.e.,

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

in which $P_{i,base}$ represents "a possibility of a base in a site i, considering maternal haplotype and fetal haplotype (state j, $j=\{x_i, y_i\}$)", i.e.,

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

in which, an indicator function is

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

Such step is to perform HMM parameter, calculating a probability distribution of observation in each site $b_{i,j}(s_i)$, i.e., calculating a possibility presenting current sequencing data (observations) in the pregnant plasma, assuming different fetal haplotypes in each site.

Step 3, constructing a partial probability matrix, and a reversal cursor (taking an example of constructing a one-dimensional probability matrix):

Definition: partial probability
$$\delta_i(q_i) = \left(\max_{q_{i-1} \in Q} \delta_i(q_i) \cdot a_{q_{i-1} q_i}\right) \cdot b_{i,q_i}(s_i),$$

Definition: reversal cursor
$$\Psi_i(q_i) = \arg\max_{q_{i-1} \in Q} \delta_i(q_i) \cdot a_{q_{i-1} q_i}.$$

Terms of "partial probability $\delta_i(q_i)$" and "reversal cursor $\Psi_i(q_i)$" used herein both follow classic definitions of Viterbi algorithm. Detailed descriptions for the definition of the parameter may refer to Lawrence R. Rabiner, PROCEEDINGS OF THE IEEE, Vol.77, No.2, February 1989, which is incorporated herein by reference.

Step 4, determining a final state, and tracing back an optional path

Determination of the final state, $q_N^* = \arg\max_{q_N \in Q} \delta_N(q_N)$.

The most possible fetal haplotype $q_i^* = \Psi_i(q_i)$ $(i = 1,2,3,...,N-1)$ is obtained by tracing back the optional path based on the reversal curse.

Step 5, outputting a result

Thus, the sequence of the fetal genome may be effectively analyzed. Comparing to other existing method of antenatal detection, the method of the present disclosure may have following technical advantages, mainly embodying in accuracy and amount of genetic information obtainable:

1) According to embodiments of the present disclosure, a site to be detected is not limited to a parental site, for a maternal site, i.e., a maternal heterozygous site, whether a fetus inherits a maternal pathopoiesia site may also be detected excellently, with an accuracy up to 95% or more; and a plurality of abnormality types can be detected, which enlarges a range of disease detection.

2) According to embodiments of the present disclosure, information of a plurality of site and diseases may be obtained by one time of sequencing; while those gene sequence, having a low coverage in the pregnant plasma which is not able to be accurately determined only by enhancing sequencing depth, may be obtained by the method of the present disclosure, with an accurate and liable result.

3. According to embodiments of the present disclosure, a plotting with a genetic disease may be performed, some related diseases may be directly deduced with information of other sites, with a large amount of information obtained for one time, which has a more instructive meaning for clinical detection.

In addition, according to embodiments of the present disclosure, the method of

determining base information of a predetermined region in a fetal genome, not limited to a certain genetic polymorphic sites such as SNP or STR, is adapted for all genetic polymorphic sites, which may be parallel used for a plurality of sites, to verify each other. Besides applying to antenatal noninvasive detect genomic information of a fetus, achieving a purpose of disease detection, the method of the present disclosure may also be used in noninvasive antenatal paternity identification, i.e., determining an identity of a fetus' father prior birth, providing assistance for disputes involving rearing responsibilities and obligations, property and sexual assault cases, etc.

System for determining base information of a predetermined region in a fetal genome

In another aspect of the present disclosure, there is provided a system for determining base information of a predetermined region in a fetal genome. According to embodiments of the present disclosure, referring to Fig.2, the system 1000 may comprises: a library constructing apparatus 100, a sequencing apparatus 200 and an analyzing apparatus 400.

According to embodiments of the present disclosure, the library constructing apparatus 100 is adapted for constructing sequencing library based on a genomic DNA sample of a fetus. According to embodiments of the present disclosure, the sequencing apparatus 200 is connected to the library constructing apparatus 100, and adapted for subjecting the sequencing library to sequencing, to obtain a sequencing result of the fetus consisting of a plurality of sequencing data. According to embodiments of the present disclosure, the system 1000 may also comprise a DNA sample extracting apparatus, adapted for extracting the genomic DNA sample of the fetus from pregnant peripheral blood. Thus, the system may be adapted for noninvasive antenatal detection.

According to embodiments of the present disclosure, optionally, the system may also comprise an aligning apparatus 300. According to embodiments of the present disclosure, the aligning apparatus 300 is connected to the sequencing apparatus 200, and adapted for aligning the sequencing result of the fetus to a reference sequence, to determine sequencing result deriving from the predetermined region. According to embodiments of the present disclosure, methods and devices for sequencing are not subjected to any special restrictions, including but not limited to Chain Termination

Method (Sanger); a high-throughput sequencing method is preferred. Thus, using characteristics being high-throughput and deep sequencing of these apparatus, efficiency may be further improved, by which precise and accuracy of subsequent analysis with sequencing data, such as statistical test, may be further improved. The high-throughput sequencing method includes but not limited to a Next-Generation sequencing technology or a single sequencing technology. The Next-Generation sequencing platform (Metzker ML. Sequencing technologies-the next generation. *Nat Rev Genet*. 2010 Jan; 11(1):31-46) includes but not limited to Illumina-Solexa (GATM, HiSeq2000TM, etc), ABI-Solid and Roche-454 (pyrosequencing) sequencing platform; the single sequencing platform (technology) includes but not limited to True Single Molecule DNA sequencing of Helicos Company, single molecule real-time (SMRT^{TM}) of Pacific Biosciences Company, and nonapore sequencing technology of Oxford Nanopore Technologies (Rusk, Nicole (2009-04-01). Cheap Third-Generation Sequencing. *Nature Methods* 6 (4): 244-245), etc. With gradual development of sequencing technology, a person skilled in the art may understand other sequencing methods and apparatuses may also be used for whole genome sequencing. According to specific examples of the present disclosure, the whole genome sequencing library may be subjected to sequencing by at least one selected from Illumina-Solexa, ABI-SOLiD, Roche-454 and a single molecule sequencing apparatus. According to embodiments of the present disclosure, a type of used reference sequence may be not subjected to any special restrictions, which may be any known sequences contained a target region. According to embodiments of the present disclosure, the reference sequence may use a known human reference genome. For example, according to embodiments of the present disclosure, the human reference genome is NCBI 36.3, HG18. In addition, according to embodiments of the present disclosure, alignment methods are not subjected to any special restrictions. According to specific examples, SOAP may be used for alignment.

According to embodiments of the present disclosure, the analyzing apparatus 400 is 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.

According to embodiments of the present disclosure, in the Viterbi algorithm, 0.25 is used as a probability distribution of an initial status, re/N is used as a recombination

probability, with re being 25~30, preferably re being 25, and N being a length of the predetermined region,

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

is used as a recombination transition matrix with p_r being re/N.

According to embodiments of the present disclosure, the aligning apparatus is adapted for determining a base having the highest probability based on a formula of

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

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

Analysis with sequencing data, which is detailed descripted above, is also adapted to the system for determining base information of a predetermined region in a fetal genome, which is omitted for brevity.

Thus, using the system may effectively implement the above method of determining base information of a predetermined region in a fetal genome, which may determine nucleic acid sequence of a predetermined region in a fetal genome may be determined in virtue of the hidden Markov Model, for example using the Viterbi algorithm, and referring to genetic information of a related individual, by which a prenatal genetic detection may be effectively performed with genetic information of the fetal genome.

In addition, according to embodiments of the present disclosure, the predetermined region is a site previously determined as having a genetic polymorphism, and the genetic polymorphism is at least one selected from single nucleotide polymorphism and STR.

Terms of "connected" should be broadly understood, which may refer to a direct connection or indirect connection, as long as achieving the above functional connection.

It should note that a person skilled in the art may understand that features and advantages of the method of determining base information of a predetermined region in a fetal genome described above may also adapted to the system for determining base information of a predetermined region in a fetal genome, which are omitted for brevity.

Computer readable medium

In a further aspect of the present disclosure, there is provided a computer readable medium. According to embodiments of the present disclosure, the computer readable medium includes 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. Thus, using the computer readable medium may effectively implement the above method of determining base information of a predetermined region in a fetal genome, which may determine nucleic acid sequence of a predetermined region in a fetal genome may be determined in virtue of the hidden Markov Model, for example using the Viterbi algorithm, and referring to genetic information of a related individual, by which a prenatal genetic detection may be effectively performed with genetic information of the fetal genome.

According to embodiments of the present disclosure, the plurality of instructions are adapted for determining the base information of the predetermined region using the hidden Markov model based on Viterbi algorithm. According to embodiments of the present disclosure, in the Viterbi algorithm, 0.25 is used as a probability distribution of an initial status, re/N is used as a recombination probability, with re being 25~30, preferably re being 25, and N being a length of the predetermined region,

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

is used as a recombination transition matrix with p_r being re/N.

According to embodiments of the present disclosure, the plurality of instructions are further adapted for determining a base having the highest probability based on based on a formula of

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

wherein

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

.

Analysis with sequencing data, which is detailed descripted above, is also adapted to the computer readable medium, which is omitted for brevity.

In addition, according to embodiments of the present disclosure, the predetermined region is a site previously determined as having a genetic polymorphism, and the genetic polymorphism is at least one selected from single nucleotide polymorphism and STR.

As to the specification, "computer readable medium" may be any device adaptive for including, storing, communicating, propagating or transferring programs to be used by or in combination with the instruction execution system, device or equipment. More specific examples of the computer readable medium comprise but are not limited to: an electronic connection (an electronic device) with one or more wires, a portable computer enclosure (a magnetic device), a random access memory (RAM), a read only memory (ROM), an erasable programmable read-only memory (EPROM or a flash memory), an optical fiber device and a portable compact disk read-only memory (CDROM). In addition, the computer readable medium may even be a paper or other appropriate medium capable of printing programs thereon, this is because, for example, the paper or other appropriate medium may be optically scanned and then edited, decrypted or processed with other appropriate methods when necessary to obtain the programs in an electric manner, and then the programs may be stored in the computer memories.

It should be understood that each part of the present disclosure may be realized by the hardware, software, firmware or their combination. In the above embodiments, a plurality of steps or methods may be realized by the software or firmware stored in the memory and executed by the appropriate instruction execution system. For example, if it is realized by the hardware, likewise in another embodiment, the steps or methods may be realized by one or a combination of the following techniques known in the art: a discrete logic circuit having a logic gate circuit for realizing a logic function of a data signal, an application-specific integrated circuit having an appropriate combination logic gate circuit, a programmable gate array (PGA), a field programmable gate array (FPGA), etc.

Those skilled in the art shall understand that all or parts of the steps in the above exemplifying method of the present disclosure may be achieved by commanding the related hardware with programs. The programs may be stored in a computer readable storage medium, and the programs comprise one or a combination of the steps in the method embodiments of the present disclosure when run on a computer.

In addition, each function cell of the embodiments of the present disclosure may be

integrated in a processing module, or these cells may be separate physical existence, or two or more cells are integrated in a processing module. The integrated module may be realized in a form of hardware or in a form of software function modules. When the integrated module is realized in a form of software function module and is sold or used as a standalone product, the integrated module may be stored in a computer readable storage medium.

Reference will be made in detail to examples of the present disclosure. It would be appreciated by those skilled in the art that the following examples are explanatory, and cannot be construed to limit the scope of the present disclosure. If the specific technology or conditions are not specified in the examples, a step will be performed in accordance with the techniques or conditions described in the literature in the art (for example, referring to J. Sambrook, *et al.* (translated by Huang PT), Molecular Cloning: A Laboratory Manual, 3rd Ed., Science Press) or in accordance with the product instructions. If the manufacturers of reagents or instruments are not specified, the reagents or instruments may be commercially available, for example, from Illumina company.

General method

The method according to embodiments of the present disclosure mainly comprises following steps:

1) noninvasive sampling a pregnant sample containing fetal genetic materials, extracting genomic DNA therefrom;

2) extracting and purifying genomic DNA sample from family members of the fetus, such as parents or grandparents thereof;

3) constructing a sequencing library with every genetic material in accordance with an requirement for different sequencing platform;

4) filtering obtained sequencing data, with filtering criteria based on quality value, adaptor contamination and etc;

5) assembling obtained high-quality sequences as required, aligning an assembled result to a human genome reference sequence, to obtain uniquely-mapped sequences for analyzing using the model.

Analysis model

Marker:

I. the number of sites to be detected is N.

II. haplotypes of parents are respectively recorded as $FH = \{fh_0, fh_1\}$ and $MH = \{mh_0, mh_1\}$,

in which

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

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

$k \in \{0,1\}$, $i = 1,2,3,...,N$.

III. Unknown fetal haplotype is recorded as $H = \{h_0, h_1\}$, particularly, h_0 and h_1 respectively represent inheriting from mother and father.

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

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

Subscripts x_i and y_i respectively present sequence pairs, and $q_i = \{x_i, y_i\}$ represents the hidden states which need to be decoded.

While, all hidden states possible presenting constitutes a set Q.

IV. Sequencing data is recorded as $S = \{s_1,...,s_i,...,s_N\}$

in which $s_i = \{n_{i,A}, n_{i,C}, n_{i,G}, n_{i,G}\}$ represents sequencing information of a site, containing the number of four bases, A, C, T and G.

V. A mean fetal concentration and a mean sequencing error rate are respectively recorded as ε and e.

Step 1, constructing a probability distribution vector of an initial state and a transition matrix of haplotypes recombination:

I. The probability distribution of the initial states is recorded as $\pi = \{\pi_j\}$ $(j \in Q)$.

According to embodiments of the present disclosure, under a circumstance of having no reference data, it may assume that $\pi_j = \Pr(q_1 = j) \triangleq \frac{1}{4}$, i.e., possibilities of

each hidden state presenting at the first site are equal.

II. According to embodiments of the present disclosure, a probability of haplotype recombination is recorded as $p_r = re/N$, in which re represents a mean times of human gamete recombinations, with a prior data ranging from 25 to 30.

III. According to embodiments of the present disclosure, a transition matrix of haplotypes recombination is recorded as $A = \{a_{jk}\}$ $(j,k \in Q)$, in which a_{jk} represents a probability of hidden states transition, i.e.,

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

Subscripts x_i and y_i of fetal haplotypes $h_0 = \{m_{1,x_1},...,m_{i,x_i},...,m_{N,x_N}\}$ and $h_1 = \{f_{1,y_1},...,f_{i,y_i},...,f_{N,y_N}\}$ constitute a sequence pair, $q_i = \{x_i, y_i\}$ constitute the hidden states to be encoded. For example, $x_i = 0$ represents "in a maternal chromosome, an allele in the corresponding locus is $m_{i,0}$".

Step 2, constructing a probability matrix of observations:

According to embodiments of the present disclosure, the probability matrix of observations is recorded as $B = \{b_{i,j}(s_i)\}$ ($i = 1,2,3,...,N$, $j \in Q$), in which $b_{i,j}(s_i)$ represents "an observed possibility of this sequencing information in a site i, considering maternal haplotype and fetal haplotype (state j, j={x_i, y_i})", i.e.,

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

in which $P_{i,base}$ represents "a possibility of a base in a site i, considering maternal haplotype and fetal haplotype (state j, j={x_i, y_i})", i.e.,

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

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

in which, an indicator function is

Step 3, constructing a partial probability matrix, and a reversal cursor (taking an example of constructing a one-dimensional probability matrix):

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

Definition: partial probability ,

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

Definition: reversal cursor .

Step 4, determining a final state, and tracing back an optional path

$$q_N^*=\arg\max_{q_N\in Q}\delta_N(q_N)$$

Determination of the final state, .

The most possible fetal haplotype $q_i^*=\Psi_i(q_i)$ ($i=1,2,3,...,N-1$) is obtained by tracing back the optional path based on the reversal curse.

Step 5, outputting a result

EXAMPLE 1

Sample collection and treatment

(1) collected sample included: peripheral blood extracted from a father and a pregnant mother within a family, and fetal umbilical cord blood after birth, all of which were collected in a tube containing EDTA for anticoagulation; saliva were collected from four grandparents using a Oragene® DNA saliva collection/DNA purification kit OG-250.

(2) extracted saliva DNA of the four grandparents were subjected to genotyping using Infinium® HD Human610-Quad BeadChip gene chip.

(3) the peripheral blood collected from the pregnant mother was centrifuged with 1600 g at 4°C for 10 min, to separate blood cells and plasma. Then obtained plasma was centrifuged with 16000 g at 4°C for 10 min, to further remove residual leukocytes, to obtain final plasma of the pregnant mother. Then genomic DNA was extracted from the final plasma of the pregnant mother using TIANamp Micro DNA Kit (TIANGEN), to obtain a genomic DNA mixture of mother and fetus thereof. Then maternal genomic DNA was extracted from removed residual leukocytes. Obtained plasma DNA were subjected to library construction based on requirement for HiSeq2000™ sequencer of Illumia® sequencer. Constructed libraries were subjected to a distribution test using Agilent®Bioanalyzer 2100 to meet a requirement for fragment ranges. Then two libraries were subjected to quantification using Q-PCR method. Qualified libraries were subjected

to sequencing using Illumina® HiSeq2000™ sequencer, with a sequencing cycle of PE101index (i.e., pair-end 101 bp index sequencing), in which parameter settings and operations were based on Illumina® specifications (obtained at http://www.illumina.com/support/documentation.ilmn)

(4) parental peripheral blood, leukocytes extracted from maternal peripheral blood and fetal umbilical cord blood were extracted with their respective genomic DNA using TIANamp Micro DNA Kit (TIANGEN).

Except for plasma DNA sample, all obtained DNA sample needed to be fragmented using Covaris™ to have a length of 500 bp. Obtained DNA fragments and plasma DNA sample were subjected to library construction based on the requirement for HiSeq2000™ sequencer of Illumia® sequencer, with a detailed procedure:

End-reparing reacting system:

10× T4 Polynucleotide kinase buffer	10 μL
dNTPs (10 mM)	4 μL
T4 DNA polymerase	5 μL
Klenow fragments	1 μL
T4 Polynucleotide kinase	5 μL
DNA fragments	30 μL
ddH_2O	up to 100 μL

After reacting at 20°C for 30 min, PCR Purification Kit (QIAGEN) was used in recycling end-repaired products. Then the recycled end-repaired products were finally dissolved in 34μL of EB buffer.

A reacting system for adding base A at end:

10× Klenow buffer	5 μL
dATP (1 mM)	10 μL
Klenow (3'-5' exo^-)	3 μL
DNA	32 μL

After incubating at 37°C for 30 min, obtained products were purified by MinElute® PCR Purification Kit (QIAGEN) and dissolved in 12 μL of EB buffer, to obtain DNA samples added with base A at end.

Ligating adaptor reacting system:

2× Rapid DNA ligating buffer	25 μL
PEI Adapter oligo-mix (20 μM)	10 μL

T4 DNA ligase	5 μL
DNA sample added with base A at end	10 μL

After reacting at 20°C for 15 min, PCR Purification Kit (QIAGEN) was used in recycling ligated products. The ligated products were finally dissolved in 32 μL of EB buffer.

PCR reacting system:

Ligated product	10 μL
Phusion DNA Polymerase Mix	25 μL
PCR primer (10 pmol/μL)	1 μL
Index N (10 pmol/μL)	1 μL
UltraPureTM Water	13 μL

Reacting procedure was shown as below:

98°C	30 s	
98°C	10 s	10 cycles
65°C	30 s	
72°C	30 s	
72°C	5 min	
4°C	Hold	

PCR Purification Kit (QIAGEN) was used in recycling PCR products, which were finally dissolved in 50 μL of EB buffer.

Constructed libraries were subjected to a distribution test using Agilent®Bioanalyzer 2100 to meet a requirement for fragment ranges. Then two libraries were subjected to quantification using Q-PCR method. Qualified libraries were subjected to sequencing using Illumina® HiSeq2000™ sequencer, with a sequencing cycle of PE101index (i.e., pair-end 101 bp index sequencing), in which parameter settings and operations were based on Illumina® specifications (obtained at http://www.illumina.com/support/documentation.ilmn)

(5) parental and maternal genomes sequencing genotyping

a. the sequencing data were aligned to a human reference genome (Hg19, NCBI 36.3) using SOAP2.

b. obtained data were subjected to consensus sequence (CNS) construction using SOAPsnp (thousands of planning data were used for Southern Han (CHS) pedigree data).

c. genotypes of a maker site were extracted.

(6) determination of parents' haplotypes

a. constructing a group genotype matrix containing ancestors' and parents' genotypes, i.e., extracting genotypes in the marker site of parents, ancestors and Southern Han pedigree.

b. deducing parents' haplotypes using BEAGLE.

(7) determination of fetal haplotype

a. aligning plasma sequencing data to a human reference genome ((Hg19, NCBI 36.3) using SOAP2;

b. constructing a probability vector of initial states, and a transition matrix of haplotypes recombination,

constructing the probability vector of initial states: taking a model of non-reference data, i.e., probabilities of every initial states were equal, being 0.25.

constructing the transition matrix of haplotypes recombination: conservatively, $re = 25$ (others were same as descriptions in "general method");

c. calculating sequencing information of each site, and constructing a probability matrix of observations (others were same as descriptions in "general method");

d. constructing a partial probability matrix, and a reversal curse (others were same as descriptions in "general method");

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

f. outputting.

According to genotyping results, the accuracy thereof were shown below:

			mother						total		
			homozygosis			heterozygosis					
			site number	accurate number	accuracy	site number	accurate number	accuracy	site number	accurate number	accuracy
autosome	father	homozygosis	199,552	199,552	100.00%	66,238	63,968	96.57%	265,790	263,520	99.15%
		heterozygosis	65,409	64,735	98.97%	41,849	39,944	95.45%	107,258	104,679	97.60%
			264,961	264,287	99.75%	108,087	103,912	96.14%	373,048	368,199	98.70%
chromosome X			4,881	4,881	100.00%	1,718	1,478	86.03%	6,599	6,359	96.36%

Industrial Applicability

The method of determining base information of a predetermined region in a fetal genome, the system for determining base information of a predetermined region in a fetal genome and a computer readable medium according to embodiments of the present disclosure may be effectively applied in analyzing the nucleic acid sequence of the predetermined region in the fetal genome.

Although explanatory embodiments have been shown and described, it would be appreciated by those skilled in the art that the above embodiments cannot be construed to limit the present disclosure, and changes, alternatives, and modifications can be made in the embodiments without departing from spirit, principles and scope of the present disclosure.

Reference throughout this specification to "an embodiment," "some embodiments", "one embodiment", "another example", "an example", "a specific examples", or "some examples", means that a particular feature, structure, material, or characteristic described in connection with the embodiment or example is included in at least one embodiment or example of the present disclosure. Thus, the appearances of the phrases such as "in some embodiments," "in one embodiment", "in an embodiment", "in another example, "in an example," "in a specific examples," or "in some examples," in various places throughout this specification are not necessarily referring to the same embodiment or example of the present disclosure. Furthermore, the particular features, structures, materials, or characteristics may be combined in any suitable manner in one or more embodiments or examples.

2012380221 01 Sep 2016

WHAT IS CLAIMED IS:

1. 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.

2. The method of claim 1, wherein the genomic DNA sample of the fetus is extracted from pregnant peripheral blood.

3. The method of claim 1 or claim 2, further comprising a step of aligning the sequencing result of the fetus to a reference sequence, to determine sequencing result deriving from the predetermined region,

wherein optionally the reference sequence is a human reference genome,

wherein the step of aligning the sequencing result of the fetal genome to the reference sequence to determine sequencing result deriving from the predetermined region further comprises:

determining a base having the highest probability based on a formula of

2012380221 01 Sep 2016

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

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

4. The method of any one of claims 1 to 3, wherein the related individual is a parent of the fetus.

5. The method of any one of claims 1 to 4, wherein the step of determining the base information of the predetermined region using the hidden Markov Model is performed based on Viterbi algorithm,

wherein optionally in the Viterbi algorithm, 0.25 is used as a probability distribution of an initial status, re/N is used as a recombination probability, with re being between about 25 and about 30, preferably re being 25, and N being a length of the predetermined region,

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

is used as a recombination transition matrix with p_r being re/N.

6. The method of any one of claims 1 to 5, wherein the predetermined region is a site previously determined as having a genetic polymorphism,

wherein optionally the genetic polymorphism is at least one selected from single nucleotide polymorphism and STR.

7. 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.

8. The system of claim 7, further comprising a DNA sample extracting apparatus, adapted for extracting the genomic DNA sample of the fetus from pregnant peripheral blood.

9. The system of claim 2 or claim 8, further comprising an aligning apparatus, connected to the sequencing apparatus, and adapted for aligning the sequencing result of the fetus to a reference sequence, to determine sequencing result deriving from the predetermined region,

wherein the aligning apparatus is adapted for determining a base having the highest probability based on a formula of

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

wherein

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

2012380221 01 Sep 2016

10. The system of any one of claims 7 to 9, wherein the analyzing apparatus is adapted for determining the base information of the predetermined region using a hidden Markov Model based on Viterbi algorithm, wherein optionally in the Viterbi algorithm, 0.25 is used as a probability distribution of an initial status, re/N is used as a recombination probability, with re being between about 25 and about 30, preferably re being 25, and N being a length of the predetermined region,

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

is used as a recombination transition matrix with p_r being re/N.

11. 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.

12. The computer readable medium of claim 11, wherein the plurality of instructions are adapted for determining the base information of the predetermined region using the hidden Markov model based on Viterbi algorithm,

wherein optionally in the Viterbi algorithm, 0.25 is used as a probability distribution of an initial status, re/N is used as a recombination probability, with re being between about 25 and about 30, preferably re being 25, and N being a length of the predetermined region,

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

is used as a recombination transition matrix with p_r being re/N.

13. The computer readable medium of claim 11 or claim 12, wherein the plurality of instructions are adapted for aligning the sequencing result of the fetus to a reference sequence, to determine sequencing result deriving from the predetermined region,

wherein optionally the plurality of instructions are further adapted for determining a base having the highest probability based on based on a formula of

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

wherein

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

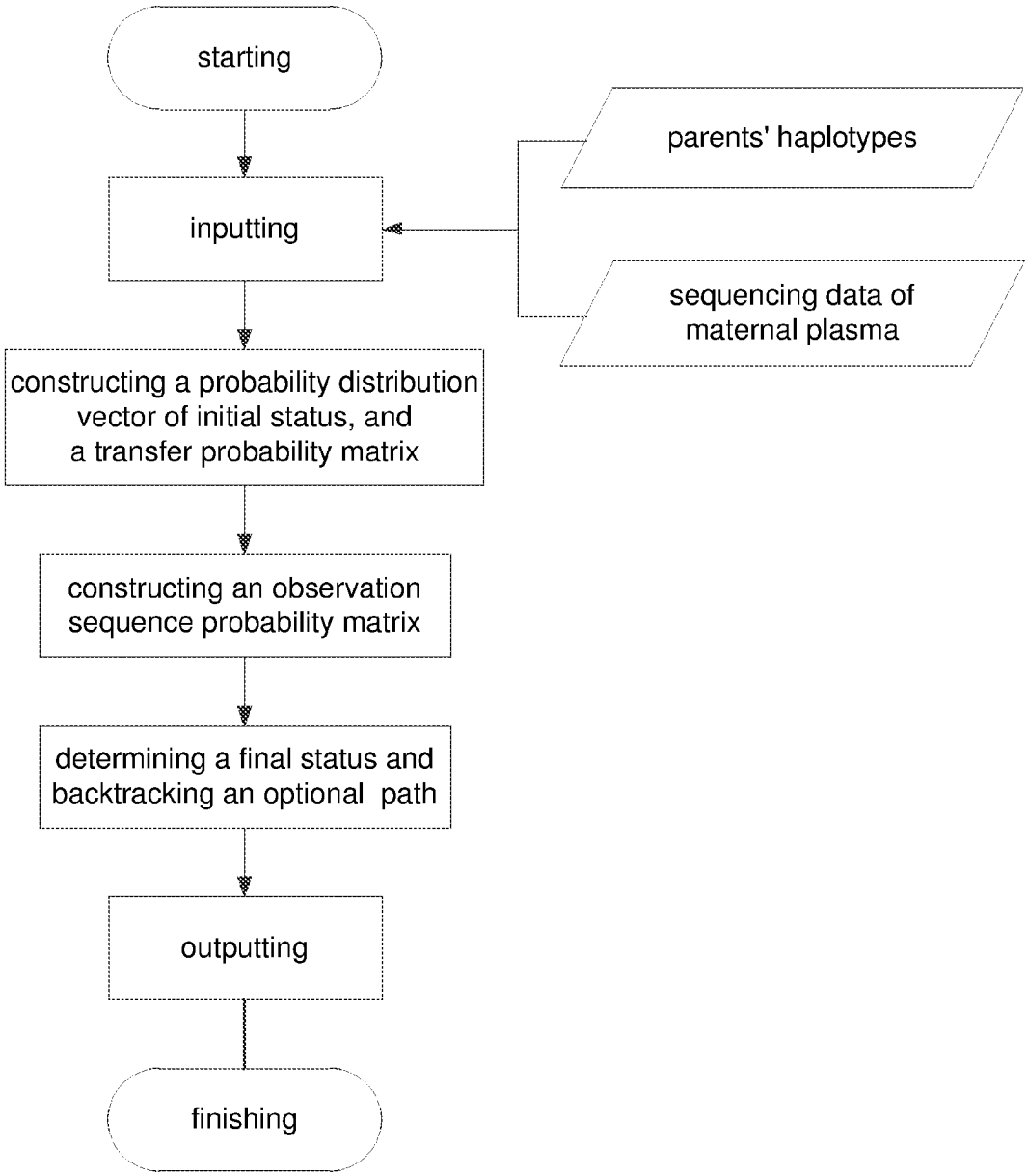

starting
inputting
parents' haplotypes
sequencing data of maternal plasma
constructing a probability distribution vector of initial status, and a transfer probability matrix
constructing an observation sequence probability matrix
determining a final status and backtracking an optional path
outputting
finishing

Fig.1

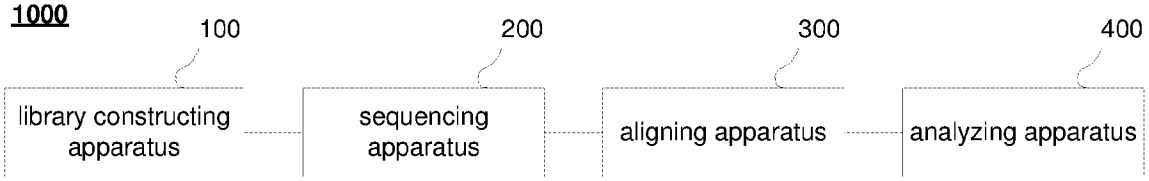

1000
100
200
300
400
library constructing apparatus
sequencing apparatus
aligning apparatus
analyzing apparatus

Fig.2

PATENT COOPERATION TREATY

PCT

INTERNATIONAL SEARCH REPORT

(PCT Article 18 and Rules 43 and 44)

Applicant's or agent's file reference	**FOR FURTHER ACTION**	see Form PCT/ISA/220 as well as, where applicable, item 5 below.
20C52575		
International application No.	International filing date *(day/month/year)*	(Earliest)Priority date *(day/month/year)*
PCT/CN2020/109602	17 Aug. 2020 (17. 08. 2020)	16 Aug. 2019 (16. 08. 2019)

Applicant

THE CHINESE UNIVERSITY OF HONG KONG

This international search report has been prepared by this International Searching Authority and is transmitted to the applicant according to Article 18. A copy is being transmitted to the International Bureau.

This international search report consists of a total of 7 sheets.

☐ It is also accompanied by a copy of each prior art document cited in this report.

1. **Basis of the report**

a. With regard to the **language**, the international search was carried out on the basis of:

☒ the international application in the language in which it was filed

☐ a translation of the international application into _____, which is the language of a translation furnished for the purposes of international search (Rules 12.3(a) and 23.1(b))

b. ☐ This international search report has been established taking into account the **rectification of an obvious mistake** authorized by or notified to this Authority under Rule 91 (Rule 43.6bis(a)).

c. ☐ With regard to any **nucleotide and /or amino acid sequence** disclosed in the international application, see Box No. I.

2. ☐ **Certain claims were found unsearchable** (see Box No. II)

3. ☐ **Unity of invention is lacking** (see Box No. III)

4. With regard to the **title**,

☒ the text is approved as submitted by the applicant.

☐ the text has been established by this Authority to read as follows:

5. With regard to the **abstract**,

☒ the text is approved as submitted by the applicant.

☐ the text has been established, according to Rule 38.2(b), by this Authority as it appears in Box IV.

The applicant may, within one month from the date of mailing of this international search report, submit comments to this Authority.

6. With regard to the **drawings**,

a. The figure of the **drawings** to be published with the abstract is Figure No.

☐ as suggested by the applicant

☐ as selected by this Authority, because the applicant failed to suggest a figure

☐ as selected by this Authority, because this figure better characterizes the invention

b. ☒ none of the figures is to be published with the abstract

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2020/109692

A. CLASSIFICATION OF SUBJECT MATTER

C12Q 1/68(2018.01)i; C12N 15/11(2006.01)i

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

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
IPC: C12Q;C12N

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

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)
CNABS, VEN, DWPI, CNKI, CNTXT, USTXT, PUBMED, BAIDU, CNKI, ISI Web of Knowledge, EMBL, Patentics, nucleotide, nucleic acid, pulse, interpulse, data, modification, detect, examine, sample, model, train, target, position, input, output, signal, methylation.

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category *	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	WO 2010068289A2 (PACIFIC BIOSCIENCES CALIFORNIA, INC.) 17 Jun.2010(17.06.2010) abstract, claims 1-99, description paragraphs [0021]-[0022] and [0047]-[0048]	1-69
A	WO 2010027484A2 (PACIFIC BIOSCIENCES OF CALIFORNIA, INC.) 11 Mar.2010(11.03.2010) claims 1-42, description paragraphs [0036]-[0037]	1-69
A	WO 2014153757A1 (BGI SHENZHEN) 02 Oct.2014(02.10.2014) the whole document	1-69
A	WO 2013188846A1 (STYLLI HARRY) 19 Dec.2013(19.12.2013) the whole document	1-69
A	CN 104053789A (BGI HEALTH SERVICE CO LTD) 17 Sep.2014(17.09.2014) the whole document	1-69
A	WO 2013170429A1 (BGI HEALTH SERVICE CO LTD et al.) 21 Nov.2013(21.11.2013) the whole document	1-69
A	US 2013230909A1 (INDUSTRIAL TECHNOLOGY RESEARCH INSTITUTE) 05 Sep.2013(05.09.2013) the whole document	1-69
A	Benjamin A.F. Structural and Mechanistic Basis for Anaerobic Ergothioneine Biosynthesis. NATURE METHODS 09 May 2010(09.05.2010) No.6 Vol.7 abstract, page 462 left column paragraph 2-page 465 left column paragraph 2	1-69

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

Form PCT/ISA /210 (2 sheet) (January 2015)

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2020/109692

* 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 30 Oct. 2020(30. 10. 2020)	Date of mailing of the international search report 19 Nov. 2020(19. 11. 2020)
Name and mailing address of the ISA/CN National Intellectual Property Administration, PRC 6, Xitucheng Rd., Jimen Bridge, Haidian District, Beijing 100088, China Facsimile No.(86-10)62019451	Authorized officer PENG, Haihang Telephone No.: 86- (10) -53961949

Form PCT/ISA /210 (3 sheet) (January 2015)

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2020/109602

Patent Documents referred in the Report	Publication Date	Patent Family	Publication Date
WO2010068289A2	17.06.2010	CA2746632A1	17.06.2010
		BRPI0922411A2	05.06.2018
		US2013029853A1	31.01.2013
		EP2370598A4	13.06.2012
		AU2009325069B2	19.03.2015
		EP2370598A2	05.10.2011
		WO2010068289A3	14.10.2010
		US2010221716A1	02.09.2010
		AU2009325069A1	21.07.2011
		US9175341B2	03.11.2015
		US2013316916A1	28.11.2013
		CA2746632C	30.06.2020
		CN102317473A	11.01.2012
		EP2370598B1	15.02.2017
WO2010027484A2	11.03.2010	AU2009288696A1	11.03.2010
		WO2010027484A9	24.03.2011
		CA2735979A1	11.03.2010
		EP2326733A2	01.06.2011
		US2010075332A1	25.03.2010
		WO2010027484A3	24.06.2010
		US8530164B2	10.09.2013
WO2014153757A1	02.10.2014	CN105051208A	11.11.2015
		HK1213945A1	15.07.2016
		CN105051208B	19.04.2017
WO2013188846A1	19.12.2013	US2018245158A1	30.08.2018
		HK1209793A1	08.04.2016
		KR20150035821A	07.04.2015
		BR112014031365A2	27.06.2017
		EP2861765A4	16.03.2016
		SG11201408385TA	29.01.2015
		EA201590027A1	29.05.2015
		SG10201610508VA	27.02.2017
		CA2876731A1	19.12.2013
		EP2861765A1	22.04.2015

		IL236181D0	29.01.2015
		US2015275298A1	01.10.2015
		US2014179805A1	26.06.2014
		JP2015522260A	06.08.2015
		MX2014015425A	14.07.2015
		EP2861765B1	23.01.2019
		CN104603289A	06.05.2015
		AU2013274002A1	05.02.2015
		IL236181A	29.01.2015
CN104053789A	17.09.2014	EP2851431B1	13.12.2017
		SG11201407515RA	30.12.2014
		KR101770460B1	22.08.2017
		AU2012380221B2	29.09.2016
		JP2015525062A	03.09.2015
		US2018320235A1	08.11.2018
		US2015094210A1	02.04.2015
		CN104053789B	10.02.2016
		PL2851431T3	30.04.2018
		AU2012380221A1	06.11.2014
		EP2851431A1	25.03.2015
		RU2014150655A	10.07.2016
		RU2597981C2	20.09.2016
		ES2656023T3	22.02.2018
		JP6045686B2	14.12.2016
		KR20140146193A	24.12.2014
		EP2851431A4	27.01.2016
		HK1196401A1	26.08.2016
		WO2013170429A1	21.11.2013
WO2013170429A1	21.11.2013	EP2851431B1	13.12.2017
		SG11201407515RA	30.12.2014
		KR101770460B1	22.08.2017
		CN104053789A	17.09.2014
		AU2012380221B2	29.09.2016
		JP2015525062A	03.09.2015
		US2018320235A1	08.11.2018
		US2015094210A1	02.04.2015

		CN104053789B	10.02.2016
		PL2851431T3	30.04.2018
		AU2012380221A1	06.11.2014
		EP2851431A1	25.03.2015
		RU2014150655A	10.07.2016
		RU2597981C2	20.09.2016
		ES2656023T3	22.02.2018
		JP6045686B2	14.12.2016
		KR20140146193A	24.12.2014
		EP2851431A4	27.01.2016
		HK1196401A1	26.08.2016
US2013230909A1	05.09.2013	JP2012507990A	05.04.2012
		AU2009311073B2	26.04.2012
		TWI385253B	11.02.2013
		US9747414B2	29.08.2017
		EP2245187A1	03.11.2010
		ES2528253T3	05.02.2015
		JP5483628B2	07.05.2014
		US2018018425A1	18.01.2018
		US2015379194A1	31.12.2015
		PL2245187T3	29.05.2015
		CN103484560B	05.11.2014
		US10515714B2	24.12.2019
		US8486630B2	16.07.2013
		EP2740806B1	08.03.2017
		CN103484560A	01.01.2014
		US9767251B2	19.09.2017
		WO2010051773A1	14.05.2010
		EP2245187B1	17.12.2014
		US2012295260A1	22.11.2012
		US2010121582A1	13.05.2010
		EP2740806A3	20.08.2014
		AU2009311073A1	14.05.2010
		EP2740806A2	11.06.2014
		CN102076871B	01.01.2014
		TW201018731A	16.05.2010

		EP2245187A4	26.06.2013
		US2014256570A1	11.09.2014
		CN102076871A	25.05.2011
		DK2245187T3	30.03.2015

PATENT COOPERATION TREATY

From the INTERNATIONAL SEARCHING AUTHORITY

To:

100098

19 A, Tower A, InDo Building, No. 48A Zhichun Road, Haidian District, Beijing, China

INSIGHT INTELLECTUAL PROPERTY LIMITED

PCT

WRITTEN OPINION OF THE INTERNATIONAL SEARCHING AUTHORITY

(PCT Rule 43 *bis*.1)

	Date of mailing (*day/month/year*) 19. Nov. 2020 (19. 11. 2020)
Applicant's or agent's file reference 20C52575	**FOR FURTHER ACTION** See paragraph 2 below

International application No.	International filing date (*day/month/year*)	Priority date (*day/month/year*)
PCT/CN2020/109602	17. Aug. 2020 (17. 08. 2020)	16. Aug. 2019 (16. 08. 2019)

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

C12Q 1/68 (2018.01) i; C12N 15/11 (2006.01) i

Applicant

THE CHINESE UNIVERSITY OF HONG KONG

1. This opinion contains indications relating to the following items:

- ☒ Box No. I Basis of the opinion
- ☐ Box No. II Priority
- ☐ Box No. III Non-establishment of opinion with regard to novelty, inventive step and industrial applicability
- ☐ Box No. IV Lack of unity of invention
- ☒ Box No. V Reasoned statement under Rule 43*bis*.1(a)(i) with regard to novelty, inventive step or industrial applicability; citations and explanations supporting such statement
- ☐ Box No. VI Certain documents cited
- ☐ Box No. VII Certain defects in the international application
- ☐ Box No. VIII Certain observations on the international application

2. **FURTHER ACTION**

If a demand for international preliminary examination is made, this opinion will be considered to be a written opinion of the International Preliminary Examining Authority ("IPEA") except that this does not apply where the applicant chooses an Authority other than this one to be the IPEA and the chosen IPEA has notified the International Bureau under Rule 66.1*bis*(b) that written opinions of this International Searching Authority will not be so considered.

If this opinion is, as provided above, considered to be a written opinion of the IPEA, the applicant is invited to submit to the IPEA a written reply together, where appropriate, with amendments, before the expiration of **3** months from the date of mailing of Form PCT/ISA/220 or before the expiration of **22** months from the priority date, whichever expires later.

For further options, see Form PCT/ISA/220.

3. For further details, see notes to Form PCT/ISA/220.

Name and mailing address of the ISA/CN	Date of completion of this opinion	Authorized officer
National Intellectual Property Administration, PRC 6, Xitucheng Rd., Jimen Bridge, Haidian District, Beijing 100088, China Facsimile No. (86-10) 62019451	16. Nov. 2020 (16. 11. 2020)	**PENG,Haihang** Telephone No. **86-(10)-53961949**

WRITTEN OPINION OF THE INTERNATIONAL SEARCHING AUTHORITY

International application No.
PCT/US2020/109602

Box No. I Basis of the opinion

1. With regard to the **language**, this opinion has been established on the basis of:
   - ☒ the international application in the language in which it was filed.
   - ☐ a translation of the international application into_____ which is the language of a translation furnished for the purposes of international search (Rules 12.3(a) and 23.1(b)).

2. ☐ This opinion has been established taking into account the **rectification of an obvious mistake** authorized by or notified to this Authority under Rule 91(Rule 43bis.1(b))

3. ☐ With regard to any **nucleotide and/or amino acid sequence** disclosed in the international application, this opinion has been established on the basis of a sequence listing:
   - a. ☐ forming part of the international application as filed:
     - ☐ in the form of an Annex C/ST.25 text file.
     - ☐ on paper or in the form of an image file.
   - b. ☐ furnished together with the international application under PCT Rule 13ter.1(a) for the purposes of international search only in the form of an Annex C/ST.25 text file.
   - c. ☐ furnished subsequent to the international filing date for the purposes of international search only:
     - ☐ in the form of an Annex C/ST.25 text file (Rule 13ter.1(a)).
     - ☐ on paper or in the form of an image file (Rule 13ter.1(b) and Administrative Instructions, Section 713).

4. ☐ In addition, in the case that more than one version or copy of a sequence listing has been filed or furnished, the required statements that the information in the subsequent or additional copies is identical to that forming part of the application as filed or does not go beyond the application as filed, as appropriate, were furnished.

5. Additional comments:

WRITTEN OPINION OF THE INTERNATIONAL SEARCHING AUTHORITY

International application No. PCT/US2020/109602

Box No. V Reasoned statement under Rule 43*bis*.1(a)(i) with regard to novelty, inventive step or industrial applicability; citations and explanations supporting such statement

1. Statement:

Novelty (N)	Claims 1-69	YES
	Claims NONE	NO
Inventive step (IS)	Claims 1-69	YES
	Claims NONE	NO
Industrial applicability (IA)	Claims 1-69	YES
	Claims NONE	NO

2. Citations and explanations

[1] **Reference are made to the following documents:**

[2] D1: WO2010/068289 A2.

[3] D2: WO2010/027484 A2.

[4] **I. Novelty**

[5] D1 discloses a method for identifying a modification, the method comprising: a) providing a template nucleic acid comprising the modification, the modification is selected from the group consisting of a methylated base, a pseudouridine base, a mismatched base, a secondary structure, or a bound agent, the template nucleic acid comprises a single-stranded portion and a double-stranded portion, meanwhile, the template nucleic acid is subjected to a treatment to alter the modification prior to the contacting, the treatment is selected from the group consisting of: glycosylase modification, bisulfite modification, DMS modification, cytosine methyltransferase modification, TETI modification, restriction enzyme modification, and cytidine deaminase 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. The change is a kinetic difference in the processing, and the kinetic difference is detected as an alteration in interpulse duration during the processing. D1 also disclose a system for identification of modifications within a nucleic acid template and a computer program product. 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. 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.(see abstract, claims 1-99, description paragraphs 0021-0022 and 0047-0048).

[6] D2 discloses 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. 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. D2 also discloses a nucleic acid sequencing system (see claims 1-42, description paragraphs [0036]-[0037]).

[7] Not all the features of claims 1-69 are disclosed in D1 or D2. Thus, these claims are new and meet the requirement of PCT Article 33(2).

[8] **II. Inventive step**

[9] D1 is regarded as the closest prior art to the subject-matter of claim 1. The subject-matter of claim 1 differs from D1 in the data structure and the model. It is not obvious to one ordinary skilled in the art at the time of invention to obtain the data structure and the model. Therefore, the subject matter of claim 1 involves an inventive step, and meets the requirement of PCT Article 33 (3).

[10] For the same reasons as above, claims 18, 46, 62, 65-69 meet the requirements of PCT Article 33(3).

[11] For similar reasons, claims 2-17, 19-45, 47-61, 63-64 meet the requirements of PCT Article 33(3).

[11] **III. Industrial applicability**

[12] The inventions of claims 1-69 can find the industrial applicability in the field of detecting diseases, thus meet the requirement of PCT Article 33(4).

Electronic Acknowledgement Receipt

EFS ID:	41974434
Application Number:	16995607
International Application Number:	
Confirmation Number:	1256
Title of Invention:	DETERMINATION OF BASE MODIFICATIONS OF NUCLEIC ACIDS
First Named Inventor/Applicant Name:	Yuk-Ming Dennis Lo
Customer Number:	20350
Filer:	David Dah-Wei Hsu/Heather Orosco
Filer Authorized By:	David Dah-Wei Hsu
Attorney Docket Number:	080015-028410US-1199353
Receipt Date:	19-FEB-2021
Filing Date:	17-AUG-2020
Time Stamp:	18:34:29
Application Type:	Utility under 35 USC 111(a)

Payment information:

Submitted with Payment	no

File Listing:

Document Number	Document Description	File Name	File Size(Bytes)/ Message Digest	Multi Part /.zip	Pages (if appl.)
1		1199353_Supplemental_IDS.pdf	167170 54e8bb4e6fbb0d87521b04aac235afeccd4b1c08	yes	4

	Multipart Description/PDF files in .zip description					
	Document Description		Start		End	
	Information Disclosure Statement (IDS) Form (SB08)		3		4	
	Transmittal Letter		1		2	
Warnings:						
Information:						
2	Foreign Reference	REF_WO2010068289A2.pdf	5867034 d2b814cfefd4e0091302966897 1669c8ac68ab19		no	103
Warnings:						
Information:						
3	Foreign Reference	REF_WO2010027484A2.pdf	6167896 6973757a825c2eb170e98feca1ed08746db0f601		no	117
Warnings:						
Information:						
4	Foreign Reference	REF_English_machine_translation_and_WO2014153757A1.pdf	2841273 6a9b2b55917dd9a6a30e14fe728fddf8af0de72f		no	47
Warnings:						
Information:						
5	Foreign Reference	REF_WO2013188846A1.pdf	9450909 0e2b751e10829ade73e44c51b4275ee266247543		no	197
Warnings:						
Information:						
6	Foreign Reference	REF_English_abstract_and_CN104053789A.pdf	134012 49cff70ef41eae5e22a4511e35e49cff342a5e8e		no	3
Warnings:						
Information:						
7	Foreign Reference	REF_English_abstract_and_CN104053789B.pdf	1322130 69101ca58399099edac123ba29e45af086f1fe56		no	22
Warnings:						

Information:					
8	Foreign Reference	REF_English_abstract_and_WO2013170429A1.pdf	1494968 09e21fb5a22e7e755964355ff3e40565319132a7	no	30
Warnings:					
Information:					
9	Foreign Reference	REF_AU2012380221B2.pdf	1508969 4c6020a9cf79a7c95c6547a51ac09e8a7105aa8d	no	34
Warnings:					
Information:					
10	Non Patent Literature	REF_ISR_and_WO_for_PCTCN2020109602_mailed_2020-11-19.pdf	2784000 f85bcce8d970229ce27274f38797abcd7a2e7b79	no	11
Warnings:					
Information:					
		Total Files Size (in bytes):	31738361		

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 a 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.

UNITED STATES PATENT AND TRADEMARK OFFICE

UNITED STATES DEPARTMENT OF COMMERCE
United States Patent and Trademark Office
Address: COMMISSIONER FOR PATENTS
P.O. Box 1450
Alexandria, Virginia 22313-1450
www.uspto.gov

APPLICATION NUMBER	FILING OR 371(C) DATE	FIRST NAMED APPLICANT	ATTY. DOCKET NO./TITLE
16/995,607	08/17/2020	Yuk-Ming Dennis Lo	080015-028410US-1199353

CONFIRMATION NO. 1256

20350
Kilpatrick Townsend & Stockton LLP - West Coast
Mailstop: IP Docketing - 22
1100 Peachtree Street
Suite 2800
Atlanta, GA 30309

PUBLICATION NOTICE

Title:DETERMINATION OF BASE MODIFICATIONS OF NUCLEIC ACIDS

Publication No.US-2021-0047679-A1
Publication Date:02/18/2021

NOTICE OF PUBLICATION OF APPLICATION

The above-identified application will be electronically published as a patent application publication pursuant to 37 CFR 1.211, et seq. The patent application publication number and publication date are set forth above.

The publication may be accessed through the USPTO's publically available Searchable Databases via the Internet at www.uspto.gov. The direct link to access the publication is currently http://www.uspto.gov/patft/.

The publication process established by the Office does not provide for mailing a copy of the publication to applicant. A copy of the publication may be obtained from the Office upon payment of the appropriate fee set forth in 37 CFR 1.19(a)(1). Orders for copies of patent application publications are handled by the USPTO's Public Records Division. The Public Records Division can be reached by telephone at (571) 272-3150 or (800) 972-6382, by facsimile at (571) 273-3250, by mail addressed to the United States Patent and Trademark Office, Public Records Division, Alexandria, VA 22313-1450 or via the Internet.

In addition, information on the status of the application, including the mailing date of Office actions and the dates of receipt of correspondence filed in the Office, may also be accessed via the Internet through the Patent Electronic Business Center at www.uspto.gov using the public side of the Patent Application Information and Retrieval (PAIR) system. The direct link to access this status information is currently https://portal.uspto.gov/pair/PublicPair. Prior to publication, such status information is confidential and may only be obtained by applicant using the private side of PAIR.

Further assistance in electronically accessing the publication, or about PAIR, is available by calling the Patent Electronic Business Center at 1-866-217-9197.

Office of Data Managment, Application Assistance Unit (571) 272-4000, or (571) 272-4200, or 1-888-786-0101

UNITED STATES PATENT AND TRADEMARK OFFICE

UNITED STATES DEPARTMENT OF COMMERCE
United States Patent and Trademark Office
Address: COMMISSIONER FOR PATENTS
P.O. Box 1450
Alexandria, Virginia 22313-1450
www.uspto.gov

APPLICATION NO.	FILING DATE	FIRST NAMED INVENTOR	ATTORNEY DOCKET NO.	CONFIRMATION NO.
16/995,607	08/17/2020	Yuk-Ming Dennis Lo	080015-028410US-1199353	1256

20350 7590 12/30/2020

Kilpatrick Townsend & Stockton LLP - West Coast
Mailstop: IP Docketing - 22
1100 Peachtree Street
Suite 2800
Atlanta, GA 30309

EXAMINER
GODDARD, LAURA B

ART UNIT	PAPER NUMBER
1642	

NOTIFICATION DATE	DELIVERY MODE
12/30/2020	ELECTRONIC

Please find below and/or attached an Office communication concerning this application or proceeding.

The time period for reply, if any, is set in the attached communication.

Notice of the Office communication was sent electronically on above-indicated "Notification Date" to the following e-mail address(es):

KTSDocketing2@kilpatrick.foundationip.com
ipefiling@kilpatricktownsend.com

Office Action Summary	Application No. 16/995,607	Applicant(s) Lo et al.	
	Examiner LAURA B GODDARD	Art Unit 1642	AIA (FITF) Status Yes

-- The MAILING DATE of this communication appears on the cover sheet with the correspondence address --

Period for Reply

A SHORTENED STATUTORY PERIOD FOR REPLY IS SET TO EXPIRE 3 MONTHS FROM THE MAILING DATE OF THIS COMMUNICATION.

- Extensions of time may be available under the provisions of 37 CFR 1.136(a). In no event, however, may a reply be timely filed after SIX (6) MONTHS from the mailing date of this communication.
- If NO period for reply is specified above, the maximum statutory period will apply and will expire SIX (6) MONTHS from the mailing date of this communication.
- Failure to reply within the set or extended period for reply will, by statute, cause the application to become ABANDONED (35 U.S.C. § 133). Any reply received by the Office later than three months after the mailing date of this communication, even if timely filed, may reduce any earned patent term adjustment. See 37 CFR 1.704(b).

Status

1)☑ Responsive to communication(s) filed on 8/17/2020.

☐ A declaration(s)/affidavit(s) under **37 CFR 1.130(b)** was/were filed on _____.

2a)☐ This action is **FINAL.** 2b) ☑ This action is non-final.

3)☐ An election was made by the applicant in response to a restriction requirement set forth during the interview on _____; the restriction requirement and election have been incorporated into this action.

4)☐ Since this application is in condition for allowance except for formal matters, prosecution as to the merits is closed in accordance with the practice under *Ex parte Quayle*, 1935 C.D. 11, 453 O.G. 213.

Disposition of Claims*

5) ☑ Claim(s) 1-17 and 70-71 is/are pending in the application.

5a) Of the above claim(s) _____ is/are withdrawn from consideration.

6) ☐ Claim(s) _____ is/are allowed.

7) ☑ Claim(s) 1-17 and 70-71 is/are rejected.

8) ☐ Claim(s) _____ is/are objected to.

9) ☐ Claim(s) _____ are subject to restriction and/or election requirement

* If any claims have been determined allowable, you may be eligible to benefit from the **Patent Prosecution Highway** program at a participating intellectual property office for the corresponding application. For more information, please see http://www.uspto.gov/patents/init_events/pph/index.jsp or send an inquiry to **PPHfeedback@uspto.gov.**

Application Papers

10)☐ The specification is objected to by the Examiner.

11)☐ The drawing(s) filed on _____ is/are: a)☐ accepted or b)☐ objected to by the Examiner.

Applicant may not request that any objection to the drawing(s) be held in abeyance. See 37 CFR 1.85(a).

Replacement drawing sheet(s) including the correction is required if the drawing(s) is objected to. See 37 CFR 1.121(d).

Priority under 35 U.S.C. § 119

12)☐ Acknowledgment is made of a claim for foreign priority under 35 U.S.C. § 119(a)-(d) or (f).

Certified copies:

a)☐ All b)☐ Some** c)☐ None of the:

1.☐ Certified copies of the priority documents have been received.

2.☐ Certified copies of the priority documents have been received in Application No. _____.

3.☐ Copies of the certified copies of the priority documents have been received in this National Stage application from the International Bureau (PCT Rule 17.2(a)).

** See the attached detailed Office action for a list of the certified copies not received.

Attachment(s)

1) ☑ Notice of References Cited (PTO-892)

2) ☑ Information Disclosure Statement(s) (PTO/SB/08a and/or PTO/SB/08b) Paper No(s)/Mail Date _____.

3) ☐ Interview Summary (PTO-413) Paper No(s)/Mail Date _____.

4) ☐ Other: _____.

DETAILED ACTION

Notice of Pre-AIA or AIA Status

The present application, filed on or after March 16, 2013, is being examined under the first inventor to file provisions of the AIA.

1. Claims 1-17, 70, and 71 are pending and being examined.

Claim Rejections - 35 USC § 112

The following is a quotation of 35 U.S.C. 112(b):
(b) CONCLUSION.—The specification shall conclude with one or more claims particularly pointing out and distinctly claiming the subject matter which the inventor or a joint inventor regards as the invention.

The following is a quotation of 35 U.S.C. 112 (pre-AIA), second paragraph:
The specification shall conclude with one or more claims particularly pointing out and distinctly claiming the subject matter which the applicant regards as his invention.

2. Claims 3-7 are rejected under 35 U.S.C. 112(b) or 35 U.S.C. 112 (pre-AIA), second paragraph, as being indefinite for failing to particularly point out and distinctly claim the subject matter which the inventor or a joint inventor (or for applications subject to pre-AIA 35 U.S.C. 112, the applicant), regards as the invention.

Claim 3 recites: "determining a classification of a disorder **using the presence of the modification at one or more nucleotides**."

Claims 4 and 5 recite that the disorder is cancer or determining the subject has the disorder and treating the subject for the disorder.

Claim 6 recites: "determining the classification of the disorder **uses the number of modification or the sites of the modifications**."

Claim 7 recites: "further comprising: determining a clinically-relevant DNA fraction, a fetal methylation profile, a maternal methylation profile, a presence of an imprinting gene region, or a tissue of origin **using the presence of the modification at one or more nucleotides**."

Claims 3-7 provide for the use of the presence of the modification at one or more nucleotides or use of the number of modification or the sites of the modifications, but, since the claim does not set forth any steps involved in the method/process of how they are "used", it is unclear what method/process applicant is intending to encompass. A claim is indefinite where it merely recites a use without any active, positive steps delimiting how this use is actually practiced.

How are the presence of modifications or the number of modifications used to classify a disorder or cancer? Given the method is unclear as to how the presence of modifications, what number of modifications, and what type of modifications classify what disorder or cancer, the scope of subjects classified as having a disorder or cancer and the scope of subjects treated are unclear.

How is the presence of the modification at one or more nucleotides used to determine any of a clinically-relevant DNA fraction, a fetal methylation profile, a maternal methylation profile, a presence of an imprinting gene region, or a tissue of origin? Given the method is unclear as to how the presence of the modification at one or more nucleotides, what number of modifications, and what type of modifications determine a what clinically-relevant DNA fraction, what fetal methylation profile, what maternal methylation profile, the presence of what imprinting gene region, or which tissue of origin, the claimed scope of the determined fractions, profiles, regions, and

origin of tissues is unclear. Given the above reasons, the metes and bounds of the claims cannot be determined.

3. Claims 8 and 9 are rejected under 35 U.S.C. 112(b) or 35 U.S.C. 112 (pre-AIA), second paragraph, as being indefinite for failing to particularly point out and distinctly claim the subject matter which the inventor or a joint inventor (or for applications subject to pre-AIA 35 U.S.C. 112, the applicant), regards as the invention.

Claim 8 recites: "wherein each sample of nucleic acid molecule of the plurality of sample nucleic acid molecules **has a size greater than a cutoff size**." What size does that encompass?

Claim 9 recites: "the number of sample nucleic acid molecules is **greater than a cutoff number**." What number does that encompass?

The claims are unclear with regards to what the cutoff size or cutoff number is. No cutoff is provided for the methods, therefore the scope of the size or number values encompassed by the claimed methods is unclear. Given the above reasons, the metes and bounds of the claims cannot be determined.

Claim Rejections - 35 USC § 101

35 U.S.C. 101 reads as follows:

> Whoever invents or discovers any new and useful process, machine, manufacture, or composition of matter, or any new and useful improvement thereof, may obtain a patent therefor, subject to the conditions and requirements of this title.

4. Claims 1-17, 70, and 71 are rejected under 35 U.S.C. 101 because the claimed invention is directed to an abstract idea and law of nature without significantly more. The claim(s) recite(s) detecting a modification of a nucleotide in a nucleic acid molecule by:

- **receiving an input data structure presented as a window displaying nucleotides sequenced** in a sample nucleic acid molecule, wherein the sample nucleic acid molecule is sequenced by measuring pulses in an optical signal corresponding to the nucleotides, the input data structure comprising values for the following properties: for each nucleotide within the window: an identity of the nucleotide, a position of the nucleotide with respect to a target position within the respective window, a width of the pulse corresponding to the nucleotide, and an interpulse duration representing a time between the pulse corresponding to the nucleotide and a pulse corresponding to a neighboring nucleotide;
- **inputting the input data structure into a model**, the model trained by: receiving a first plurality of first data structures, each first data structure of the first plurality of data structures corresponding to a respective window of nucleotides sequenced in a respective nucleic acid molecule of a plurality of first nucleic acid molecules, wherein each of the first nucleic acid molecules is sequenced by measuring pulses in the optical signal corresponding to the nucleotides, wherein the modification has a known first state in a nucleotide at a target position in each window of each first nucleic acid molecule, each first data structure comprising values for the same properties as the input data structure,
- **storing a plurality of first training samples**, each including one of the first

plurality of first data structures and a first label indicating the first state of the nucleotide at the target position, and

- **optimizing**, **using the plurality of first training samples**, **parameters of the model based on outputs of the model matching or not matching corresponding labels** of the first labels when the first plurality of first data structures is input to the model, wherein an output of the model specifies whether the nucleotide at the target position in the respective window has the modification,
- **determining, using the model, whether the modification is present in a nucleotide at the target position** within the window in the input data structure.

Thus, the claims are drawn to the abstract idea of mathematically manipulating and comparing previously obtained optical pulse sequencing data (for example, obtained by single-molecule, real-time (SMRT) sequencing) and displaying it in a window (presumably on a computer screen). This judicial exception is not integrated into a practical application because it is simply an abstract idea implemented on a computer. The claim(s) does/do not include additional elements that are sufficient to amount to significantly more than the judicial exception because window display of numerical values, mathematical manipulation, storage of data, and visual comparison of the sequencing data values are well-understood, routine, conventional computer functions as recognized by the court decisions listed in MPEP § 2106.05(d).

With regards to claims 3-7, the claims are drawn to a judicial exception comprising a law of nature or natural phenomenon where the presence or number of

naturally occurring nucleotide modifications are correlated to a disorder such as cancer, a profile, gene region, or tissue of origin. This judicial exception is not integrated into a practical application because the claims recite only an observation of the naturally occurring phenomenon/law of nature of nucleic acid modification data without applying the data to a practical application. The claim(s) does/do not include additional elements that are sufficient to amount to significantly more than the judicial exception because the claims recite only the abstract idea of previously obtained nucleic acid sequence data mathematically manipulated, compared and displayed in a window.

With regards to the limitation in claim 5 for a step of "treating the subject for the disorder," the judicial exception is not integrated into a practical application because this step is a generic instruction to apply the judicial exception. Generic treatment steps are no more than appending conventional steps specified at a high level of generality, are a generic direction to "apply it," and fail to supply inventive concept to the judicial exception. The step of "treating the subject for the disorder" does not apply or use the judicial exception to effect a particular treatment for a particular disorder. The treatment step of claim 5 fails to amount to an element significantly more than the judicial exception.

To obviate the rejection, there must be at least one additional element or physical step that applies, relies on, or uses the natural principle or abstract idea so that the claim amounts to significantly more than the judicial exception or abstract idea itself. The claimed method currently fails to provide a practical application of the judicial exception and abstract idea and fails to add any elements that amount to significantly more than the judicial exception and abstract idea.

Claim Rejections - 35 USC § 102

The following is a quotation of the appropriate paragraphs of 35 U.S.C. 102 that form the basis for the rejections under this section made in this Office action:

> A person shall be entitled to a patent unless –
>
> (a)(1) the claimed invention was patented, described in a printed publication, or in public use, on sale, or otherwise available to the public before the effective filing date of the claimed invention.

5. Claim(s) 1, 10, 12-14, 70 and 71 are rejected under 35 U.S.C. 102(a)(1) as being anticipated by Flusberg et al (Nature Methods, 2010, 7:461-465).

Flusberg et al teach detecting modifications of a nucleotide in a nucleic acid molecule comprising:

(a) performing single-molecule, real-time (SMRT) sequencing of a sample of nucleic acid molecule and displaying in a window (i.e. a computer screen or piece of paper) the data for:

- identity of the nucleotide T, C, G, and A in a sequence comprising at least 8 consecutive nucleotides (Figure 1);
- the position of the nucleotide;
- the width of pulse corresponding to each nucleotide;
- the interpulse duration (IPD) representing a time between the pulse corresponding to the nucleotide and a pulse corresponding to a neighboring nucleotide;

(b) inputting this data into a model trained by data corresponding to multiple nucleic acids comprising known methylated nucleotides and control nucleic acids

comprising no methylated nucleotides; wherein the data would necessarily have to be stored in order to be manipulated on a computer and displayed as shown by Flusberg et al;

(c) optimizing the model to provide an accurate correlation of data value to presence of a methylated nucleotide in a sequence, such as providing IPD ratios and pulse width (PW) ratios for the nucleotides and identifying IPDs and PWs that indicated the presence and location of methylated nucleotides (see Materials and Methods; RESULTS on p. 462-465; Table 1; Figures 2-5);

(d) using the model, determining the presence of the methylated nucleotide at a specific nucleotide in a sequence, displayed in the window (see Materials and Methods and Supplementary Figures 1, 3, 5, 6 and Supp Table 1);

wherein SMRT sequencing reads optical fluorescence signals from a dye-labeled nucleotide (abstract);

wherein the window displays two strands of nucleotides and their strand properties (Figure 1);

wherein the tested nucleic acid is circular DNA formed by ligating a hairpin adaptor on the end of the cut dsDNA molecule (p. 463, col. 2 to p. 464; Figure 4);

wherein the methylation is detected by circular consensus sequencing (p. 463, col. 2 to p. 464).

Flusberg et al explain how SMRT sequencing data reveals nucleotide modifications (p. 462, col. 1):

“Fluorescence pulses in SMRT sequencing are characterized not only by their emission spectra but also by their duration and by the interval between successive

pulses. These metrics, defined here as pulse width and interpulse duration (IPD), add valuable information about DNA polymerase kinetics. Pulse width is a function of all kinetic steps after nucleotide binding and up to fluorophore release, and IPD is determined by the kinetics of nucleotide binding and polymerase translocation. We have previously demonstrated that SMRT sequencing polymerase synthesis rates are sensitive to DNA primary and secondary structure. Therefore, we hypothesized that methylated bases in a DNA template might be detected directly on the principle that their presence affects polymerase kinetics during SMRT sequencing (Fig. 1)."

Claim Rejections - 35 USC § 103

The following is a quotation of 35 U.S.C. 103 which forms the basis for all obviousness rejections set forth in this Office action:

> A patent for a claimed invention may not be obtained, notwithstanding that the claimed invention is not identically disclosed as set forth in section 102, if the differences between the claimed invention and the prior art are such that the claimed invention as a whole would have been obvious before the effective filing date of the claimed invention to a person having ordinary skill in the art to which the claimed invention pertains. Patentability shall not be negated by the manner in which the invention was made.

6. Claim 1, 2, 8-17, 70, and 71 are rejected under 35 U.S.C. 103 as being unpatentable over US Patent 9,175,338, Flusberg et al, issued 2015.

Flusberg et al teach detecting modifications of a nucleotide in a nucleic acid molecule comprising:

(a) performing single-molecule, real-time (SMRT) sequencing of a sample of nucleic acid molecule and obtaining the data for:

- identity of the nucleotide T, C, G, and A in a sequence comprising at least 4 consecutive nucleotides;
- the position of the nucleotide;
- the width of pulse corresponding to each nucleotide;
- the interpulse duration (IPD) representing a time between the pulse corresponding to the nucleotide and a pulse corresponding to a neighboring nucleotide;

(b) inputting this data into a model trained by data corresponding to multiple nucleic acids comprising known modified or methylated nucleotides and control nucleic acids comprising no modified nucleotides; wherein the data is stored on a computer; wherein the model uses machine learning, principal component analysis, neural network, or logistic regression; (col. 11, line 39 to col. 12, line 3; col. 17, line 35 to col. 24, line 45; col. 61-73; Figure 14);

(c) optimizing the model to provide an accurate correlation of data value to presence of a modified or methylated nucleotide in a sequence, such a providing IPD ratios and pulse width (PW) ratios for the nucleotides and identifying IPDs and PWs that indicated the presence and location of methylated nucleotides (col. 61-73; claims 1-15);

(d) using the model, determining the presence, amount, or location of the modified or methylated nucleotide at a specific nucleotide in a sequence (col. 3, lines 45 to col. 5, line 57; col. 61-73; col. 75-80; claims 1-15);

wherein SMRT sequencing reads optical fluorescence signals form a dye-labeled nucleotide (col. 17, line 35 to col. 24, line 45; col. 20, lines 49-67; Figure 4);

wherein a window displays two strands of nucleotides and their strand properties (Figure 2);

wherein the tested nucleic acid is circular DNA and formed by ligating a hairpin adaptor on the end of the cut dsDNA molecule (Figure 7 and 11; col. 4, lines 44-61; col. 6, lines 54-61; col. 7, lines 39-49; col. 9, lines 1-5; col. 25, lines 1-10; col. 53, lines 32-57; col. 60, lines 45-67; col. 75, lines 31 to col. 76, line 64; col. 79, lines 19-47);

wherein the methylation is detected by circular consensus sequencing (col. 76, lines 10-21; Figure 11);

wherein the nucleic acid is sequenced (col. 11, lines 1-39; col. 60, lines 45-67);

wherein the tested nucleic acids are obtained from a biological sample of a subject (col. 2, lines 11-23; col. 29, lines 20-35; col. 39, lines 20-48; col. 48, line 61 to col. 49, line 5);

wherein the nucleotide tested is enriched or filtered, wherein the nucleotide is filtered by selecting dsDNA molecules having a size within a size range (col. 11, lines 1-25; col. 49, lines 10-15; col. 59, lines 45-67; col. 60, lines 5-67);

wherein the modified nucleotides are determined without using a circular consensus sequence or reference genome sequence alignment (col. 74, line 56 to col. 75, line 30).

wherein the tested nucleic acid aligns to a plurality of genomic regions and that a number/barcode is assigned to each genomic region (Figure 4; col. 3, lines 40-43; col. 11, lines 20-39; col. 20, lines 27-48; col. 30, lines 31-62; col. 60, lines 50-67).

Flusberg et al do not teach or demonstrate all of the sequence nucleotide positions, modifications, IPD, PW, and strand property data is displayed in one window

for comparisons, although Flusberg et al explicitly suggest their data can be graphically electronically displayed on a computer or printed out (col. 73, lines 25-35; Figure 10; col. 73, line 55 to col. 74, line 19), Flusberg et al suggest analyzing the IPDs over a window of about 20, 30, 50, 70, or 100 bases, over a fixed length, or over substantially all of the template (col. 27, lines 5-23; col. 63, lines 31-50), and Flusberg et al demonstrate graphically displaying nucleotides in their positions, modifications, IPD, PW, and strand property (Figures 1, 2, 8, 11-25).

It would have been *prima facie* obvious to one of ordinary skill in the art at the time the invention was filed to display the SMRT sequencing data of Flusberg et al in a window. One would have been motivated to and have a reasonable expectation of success because Flusberg et al suggest graphically displaying their data and successfully demonstrate graphically presenting their data.

7. Claims 3-6 are rejected under 35 U.S.C. 103 as being unpatentable over US Patent 9,175,338, Flusberg et al, issued 2015 as applied to claims 1, 2, 8-17, 70, and 71 above, and further in view of Sugai et al (Clinical Epigtenetics, 2017, 9:55, internet pages 1-10).

Flusberg et al teach a method as set forth above. Flusberg et al further teach it is known that modifications to nucleotide sequences are associated with diseases, wherein DNA methylation is known to be associated with cancer and other disease processes (col. 1, lines 25-65), and DNA damage is known to be associated with cancer, cardiovascular disease and nervous system diseases (col. 16, lines 39-60).

Flusberg et al further teach it is known that modifications to nucleotide sequences are associated with different tissue types (col. 2, lines 11-24).

Flusberg et al do not teach correlating the number or locations methylated nucleotides to a disorder such as cancer and treating the subject.

Sugai et al teach that increased methylation levels on specific genes is correlated to an increased risk of colorectal cancer. Sugai et al teach analyzing methylation levels from tumor tissue samples colorectal cancer patient who underwent surgery (Methods).

It would have been *prima facie* obvious to one of ordinary skill in the art at the time the invention was filed to correlate the presence and increased amount of nucleic acid methylation modifications to the presence of colorectal cancer in the method of Flusberg et al. One would have been motivated and have a reasonable expectation of success to classify a patient as having cancer based on the presence and increased amount of methylated nucleotides, given Flusberg et al recognize that methylation is associated with the presence of cancer, and given Sugai et al successfully demonstrate the correlation between the presence of colorectal cancer and increased methylation of nucleotides overall and particularly in specific genes. One would have been motivated to and have a reasonable expectation of success to treat a patient classified as having colorectal cancer given treatment of cancer is known and routine and given Sugai teach the colorectal cancer patients were treated with surgery.

8. Claim 7 is rejected under 35 U.S.C. 103 as being unpatentable over US Patent 9,175,338, Flusberg et al, issued 2015 as applied to claims 1, 2, 8-17, 70, and 71 above, and further in view of Kile et al (PLoS ONE, 2010, 5:e13730, internet pages 1-9).

Flusberg et al teach a method as set forth above. Flusberg et al further teach it is known that modifications to nucleotide sequences are associated with diseases, wherein DNA methylation is known to be associated with cancer and other disease processes (col. 1, lines 25-65), and DNA damage is known to be associated with cancer, cardiovascular disease and nervous system diseases (col. 16, lines 39-60). Flusberg et al further teach it is known that modifications to nucleotide sequences are associated with different tissue types (col. 2, lines 11-24).

Flusberg et al do not teach determining methylation for determining maternal or fetal methylation profiles.

Kile et al teach determining and comparing methylation profiles between mother and infant pairs to assess whether parental DNA methylation patterns influence the phenotype of their offspring. The heritability of epigenetic marks between generations is frequently used to explain the etiology of traits and diseases that do not follow Mendelian inheritance patterns (abstract; p. 1, col. 1-2).

It would have been *prima facie* obvious to one of ordinary skill in the art at the time the invention was filed to utilize the method of Flusberg to determine a maternal or fetal methylation profile. One would have been motivated to and have a reasonable expectation of success, given Kile et al teach there is motivation and reasonable expectation of success to analyze methylation levels in maternal and fetal pairs.

9. **Conclusion:** No claim is allowed.

10. Any inquiry concerning this communication or earlier communications from the examiner should be directed to LAURA B GODDARD whose telephone number is (571)272-8788. The examiner can normally be reached on Mon-Fri, 7am-3:30pm.

Examiner interviews are available via telephone, in-person, and video conferencing using a USPTO supplied web-based collaboration tool. To schedule an interview, applicant is encouraged to use the USPTO Automated Interview Request (AIR) at http://www.uspto.gov/interviewpractice.

If attempts to reach the examiner by telephone are unsuccessful, the examiner's supervisor, Misook Yu can be reached on 571-272-0839. The fax phone number for the organization where this application or proceeding is assigned is 571-273-8300.

Information regarding the status of an application may be obtained from the Patent Application Information Retrieval (PAIR) system. Status information for published applications may be obtained from either Private PAIR or Public PAIR. Status information for unpublished applications is available through Private PAIR only. For more information about the PAIR system, see https://ppair-my.uspto.gov/pair/PrivatePair. Should you have questions on access to the Private PAIR system, contact the Electronic Business Center (EBC) at 866-217-9197 (toll-free). If you would like assistance from a USPTO Customer Service Representative or access

to the automated information system, call 800-786-9199 (IN USA OR CANADA) or 571-272-1000.

/Laura B Goddard/
Primary Examiner, Art Unit 1642

Notice of References Cited	Application/Control No. 16/995,607	Applicant(s)/Patent Under Reexamination Lo et al.	
	Examiner LAURA B GODDARD	Art Unit 1642	Page 1 of 1

U.S. PATENT DOCUMENTS

*		Document Number Country Code-Number-Kind Code	Date MM-YYYY	Name	CPC Classification	US Classification
*	A	US-9175338-B2	11-2015	Flusberg; Benjamin	C12Q1/6869	1/1
	B					
	C					
	D					
	E					
	F					
	G					
	H					
	I					
	J					
	K					
	L					
	M					

FOREIGN PATENT DOCUMENTS

*		Document Number Country Code-Number-Kind Code	Date MM-YYYY	Country	Name	CPC Classification
	N					
	O					
	P					
	Q					
	R					
	S					
	T					

NON-PATENT DOCUMENTS

*		Include as applicable: Author, Title Date, Publisher, Edition or Volume, Pertinent Pages)
	U	Flusberg et al (Nature Methods, 2010, 7:461-465)
	V	Sugai et al (Clinical Epigtenetics, 2017, 9:55, internet pages 1-10)
	W	Kile et al (PLoS ONE, 2010, 5:e13730, internet pages 1-9)
	X	Flusberg et al (Nature Methods, 2010, 7:461-465) Supplemental Data

*A copy of this reference is not being furnished with this Office action. (See MPEP § 707.05(a).)
Dates in MM-YYYY format are publication dates. Classifications may be US or foreign.

Search Notes	Application/Control No.	Applicant(s)/Patent Under Reexamination
	16/995,607	Lo et al.
	Examiner	Art Unit
	LAURA B GODDARD	1642

CPC - Searched*		
Symbol	Date	Examiner

CPC Combination Sets - Searched*		
Symbol	Date	Examiner

US Classification - Searched*			
Class	Subclass	Date	Examiner

* See search history printout included with this form or the SEARCH NOTES box below to determine the scope of the search.

Search Notes		
Search Notes	Date	Examiner
WEST	12/23/2020	LBG
Google	12/23/2020	LBG
STN	12/23/2020	LBG

Interference Search			
US Class/CPC Symbol	US Subclass/CPC Group	Date	Examiner

/Laura B Goddard/ Primary Examiner, Art Unit 1642	

Web Search History

date, time	web site	search string
12/22/2020 2:53:52 PM	Google	optical signal pulse sequencing nuclkeotide
12/22/2020 2:54:12 PM	Google	optical detection dsDNA ssDNA compare methylation
12/22/2020 2:54:57 PM	Google	Single-Molecule DNA Methylation Quantification Using Electro-Optical Sensing in Solid-State Nanopores
12/22/2020 3:02:16 PM	Google	SMRT sequencing
12/22/2020 3:02:40 PM	Google	"Single-molecule real-time sequencing" compare methylated unmethylated
12/22/2020 4:10:15 PM	Google	De novo identification of DNA modifications enabled by genome-guided Nanopore signal processing.
12/22/2020 6:39:38 PM	Google	cas9 AND SMRT sequencing
12/22/2020 6:40:40 PM	Google	CRISPR guided SMRT sequencing circular
12/22/2020 6:49:39 PM	Google	cas9 AND SMRT sequencing circular
12/22/2020 6:53:26 PM	Google	. Loose M, Malla S, Stout M (2016) Real-time selective sequencing using nanopore technology. NatMethod
12/22/2020 6:59:35 PM	Google	Targeted nanopore sequencing with Cas9 for studies of methylation, structural variants andmutations
12/23/2020 10:31:37 AM	Google	Direct detection of DNA methylation during single-molecule, real-time sequencing
12/23/2020 10:57:09 AM	Google	logistic regression
12/23/2020 10:57:13 AM	Google	logistic regression classification-and-regression trees (CART)
12/23/2020 1:21:11 PM	Google	methylation levels
12/23/2020 1:21:44 PM	Google	methylation levels correlate cancer tissues of origin
12/23/2020 1:32:22 PM	Google	methylation levels correlate cancer
12/23/2020 1:32:55 PM	Google	methylation levels correlate amternal methylation

INFORMATION DISCLOSURE STATEMENT BY APPLICANT (Not for submission under 37 CFR 1.99)		
	Application Number	16/995,607
	Filing Date	August 17, 2020
	First Named Inventor	Yuk-Ming Dennis Lo
	Art Unit	Not Yet Known
	Examiner Name	Not Yet Known
	Attorney Docket Number	080015-028410US-1199353

U.S. PATENTS

Examiner Initial*	Cite No	Patent Number	Kind Code [1]	Issue Date	Name of Patentee or Applicant of cited Document	Pages, Columns, Lines, Where Relevant Passages or Relevant Figures Appear
	1.	9,116,118	B2	08-25-2015	Turner et al.	
	2.	9,175,338	B2	11-03-2015	Flusberg et al.	
	3.	10,590,484	B2	03-17-2020	Korlach et al.	
	4.	8,467,976	B2	06-18-2013	Lo et al.	
	5.	10,468,121	B2	06-18-2013	Lo et al.	
	6.	10,465,232	B1	11-05-2019	Wu et al.	
	7.	10,438,691	B2	10-08-2019	Kim et al.	
	8.	9,175,341	B2	11-03-2015	Flusberg et al.	

U.S. PATENT APPLICATION PUBLICATIONS

Examiner Initial*	Cite No	Publication Number	Kind Code [1]	Publication Date	Name of Patentee or Applicant of cited Document	Pages, Columns, Lines, Where Relevant Passages or Relevant Figures Appear
	9.	2011/0183320	A1	07-28-2011	Flusberg et al.	
	10.	2017/0037464	A1	02-09-2017	Turner et al.	
	11.	2017/0233802	A1	08-17-2017	Flusberg et al.	
	12.	2018/0201993	A1	07-19-2018	Turner et al.	
	13.	2019/0024162	A1	01-24-2019	Korlach et al.	
	14.	2017/0159132	A1	06-08-2017	Okino et al.	
	15.	2014/0004511	A1	01-02-2014	Korlach et al.	

FOREIGN PATENT DOCUMENTS

Examiner Initial*	Cite No	Foreign Document Number[3]	Country Code[2]	Kind Code[4]	Publication Date	Name of Patentee or Applicant of cited Document	Pages, Columns, Lines where Relevant Passages or Relevant Figures Appear	T[5]
	16.	101896167	KR	B1	09-07-2018		English abstract attached. Equivalent to WO 2009013492 A1 below.	
	17.	2009013492	WO	A1	01-29-2009	The Chinese University of Hong Kong	Equivalent to KR 101896167 B1 above.	

NON-PATENT LITERATURE DOCUMENTS

Examiner Initials*	Cite No	Include name of the author (in CAPITAL LETTERS), title of the article (when appropriate), title of the item (book, magazine, journal, serial, symposium, catalog, etc.), date, page(s), volume-issue number(s), publisher, city and/or country where published.	T[5]

ALL REFERENCES CONSIDERED EXCEPT WHERE LINED THROUGH. /LBG/

INFORMATION DISCLOSURE STATEMENT BY APPLICANT (Not for submission under 37 CFR 1.99)		
	Application Number	16/995,607
	Filing Date	August 17, 2020
	First Named Inventor	Yuk-Ming Dennis Lo
	Art Unit	Not Yet Known
	Examiner Name	Not Yet Known
	Attorney Docket Number	080015-028410US-1199353

	18.	FLUSBERG, Benjamin A. et al.; "Direct detection of DNA methylation during single-molecule, real-time sequencing"; Nature Methods; June 2010; Vol. 7, No. 6; pages 461-465 (7 pages).	
	19.	LIU, Qian et al.; "Detection of DNA base modifications by deep recurrent neural network on Oxford Nanopore sequencing data"; Nature Communications; 2019; 10(1):2449; doi: 10.1038/s41467-019-10168-2; 11 pages.	
	20.	NI, Peng et al.; "DeepSignal: detecting DNA methylation state from Nanopore sequencing reads using deep-learning"; Bioinformatics; 2019; Vol. 35, No. 22; doi: 10.1093/bioinformatics/btz276; 10 pages.	
	21.	SUZUKI, Yuta et al.; "AgIn: measuring the landscape of CpG methylation of individual repetitive elements"; Bioinformatics; 2016; Vol. 32, No. 19; pages 2911-2919.	
	22.	BLOW, Matthew J. et al.; "The Epigenomic Landscape of Prokaryotes"; PLOS Genetics; February 12, 2016; 12(2): e1005854; doi:10.1371/journal.pgen.1005854; 28 pages.	
	23.	CLARK, Tyson A. et al.; "Enhanced 5-methylcytosine detection in single-molecule, real-time sequencing via Tet1 oxidation"; BMC Biology; 2013; Vol. 11, No. 4; http://www.biomedcentral.com/1741-7007/11/4; 10 pages.	
	24.	CLARK, Tyson A. et al.; "Characterization of DNA methyltransferase specificities using single-molecule, real-time DNA sequencing"; Nucleic Acids Research; 2012; Vol. 40, No. 4; e29; https://doi.org/10.1093/nar/gkr1146; 12 pages.	
	25.	EID, John et al.; "Real-Time DNA Sequencing from Single Polymerase Molecules"; Science; 02 January 2009; Volume 323, Issue 5910; DOI: 10.1126/science.1162986; page 133-138 (7 pages).	
	26.	FENG, Zhixing et al.; "Detecting DNA Modifications from SMRT Sequencing Data by Modeling Sequence Context Dependence of Polymerase Kinetic"; PLOS Computational Biology; March 2013; Volume 9, Issue 3; e1002935; https://doi.org/10.1371/journal.pcbi.1002935; 10 pages.	
	27.	LIU, Yibin et al.; "Bisulfite-free direct detection of 5-methylcytosine and 5-hydroxymethylcytosine at base resolution"; Nature Biotechnology Letters; April 2019; Volume 37; pages 424-429 (11 pages).	
	28.	SCHADT, Eric E. et al.; "Modeling kinetic rate variation in third generation DNA sequencing data to detect putative modifications to DNA bases"; Genome Research; 2013; Vol. 23, No. 1; doi: 10.1101/gr.136739.111; pages 129-141.	
	29.	FENG, Zhixing et al.; "qDNAmod: a statistical model-based tool to reveal intercellular heterogeneity of DNA modification from SMRT sequencing data"; Nucleic Acids Research; 2014; Vol. 42, No. 22; doi: 10.1093/nar/gku1097; pages 13488-13499.	

ALL REFERENCES CONSIDERED EXCEPT WHERE LINED THROUGH. /LBG/

INFORMATION DISCLOSURE STATEMENT BY APPLICANT (Not for submission under 37 CFR 1.99)		
	Application Number	16/995,607
	Filing Date	August 17, 2020
	First Named Inventor	Yuk-Ming Dennis Lo
	Art Unit	Not Yet Known
	Examiner Name	Not Yet Known
	Attorney Docket Number	080015-028410US-1199353

	30.	SUZUKI, Yuta; "Observing comprehensive DNA methylomes via single molecule real-time sequencing: application to diploid and centromeric methylation"; dissertation submitted to The Graduate School of Frontier Sciences, The University of Tokyo, on December 13, 2017; 118 pages.	

EXAMINER SIGNATURE			
Examiner Signature /Laura B Goddard/		Date Considered 12/23/2020	

*EXAMINER: Initial if reference considered, whether or not citation is in conformance with MPEP 609. Draw line through a citation if not in conformance and not considered. Include copy of this form with next communication to applicant.

[1] See Kind Codes of USPTO Patent Documents at www.USPTO.GOV or MPEP 901.04. [2] Enter office that issued the document, by the two-letter code (WIPO Standard ST.3). [3] For Japanese patent documents, the indication of the year of the reign of the Emperor must precede the serial number of the patent document.
[4] Kind of document by the appropriate symbols as indicated on the document under WIPO Standard ST.16 if possible. [5] Applicant is to place a check mark here if English language translation is attached.

ALL REFERENCES CONSIDERED EXCEPT WHERE LINED THROUGH. /LBG/

```
     (FILE 'HOME' ENTERED AT 14:02:11 ON 23 DEC 2020)
                D SET
                D CLUSTER
                D CLUSTER .BIO

     FILE 'CAPLUS, EMBASE, BIOSIS, MEDLINE, SCISEARCH' ENTERED AT 14:02:33 ON
     23 DEC 2020
L1           606 SEA SPE=ON  ABB=ON  PLU=ON  ("SINGLE-MOLECULE REAL-TIME
                 SEQUENCING" OR "SMRT SEQUENCING") AND METHYLATION
L2           514 SEA SPE=ON  ABB=ON  PLU=ON  L1 AND PY<2020
L3             6 SEA SPE=ON  ABB=ON  PLU=ON  L2 AND (DISPLAY OR WINDOW)
L4             4 DUP REM L3 (2 DUPLICATES REMOVED)
                 D L4 1-4 IBIB ABS HITIND
L5            14 SEA SPE=ON  ABB=ON  PLU=ON  L2 AND (CANCER OR TUMOR OR
                 NEOPLASM OR MALIGNANT OR CARCINOMA)
L6            12 DUP REM L5 (2 DUPLICATES REMOVED)
                 D L6 1-12 IBIB ABS HITIND
```

WEST Search History

Hide Items | Restore | Clear | Close

DATE: Tuesday, December 22, 2020

Hide?	Set Name	Query	Hit Count
	DB=PGPB,USPT,USOC,EPAB,JPAB,DWPI; PLUR=YES; OP=OR		
☐	L10	L9 AND pulse[CLM]	0
☐	L9	"single molecule real time sequencing" AND methylation AND ("Chinese University")	13
☐	L8	L7 AND (methylation methylated)[AB]	11
☐	L7	L6 AND (disease cancer tumor carcinoma malignant neoplasm)[AB]	66
☐	L6	L3 AND cancer	263
☐	L5	L4 AND SMRT	23
☐	L4	L3 AND cas9	39
☐	L3	"single molecule real time sequencing" AND methylation AND algorithm AND @ay<=2019	316
☐	L2	L1 AND @ay<=2019	13
☐	L1	"single molecule real time sequencing" AND (training) AND methylation AND IPD	13

END OF SEARCH HISTORY

UNITED STATES PATENT AND TRADEMARK OFFICE

UNITED STATES DEPARTMENT OF COMMERCE
United States Patent and Trademark Office
Address: COMMISSIONER FOR PATENTS
P.O. Box 1450
Alexandria, Virginia 22313-1450
www.uspto.gov

APPLICATION NO.	FILING DATE	FIRST NAMED INVENTOR	ATTORNEY DOCKET NO.	CONFIRMATION NO.
16/995,607	08/17/2020	Yuk-Ming Dennis Lo	080015-028410US-1199353	1256

20350 7590 11/06/2020

KILPATRICK TOWNSEND & STOCKTON LLP
Mailstop: IP Docketing - 22
1100 Peachtree Street
Suite 2800
Atlanta, GA 30309

EXAMINER	
ART UNIT	PAPER NUMBER

NOTIFICATION DATE	DELIVERY MODE
11/06/2020	ELECTRONIC

Please find below and/or attached an Office communication concerning this application or proceeding.

The time period for reply, if any, is set in the attached communication.

Notice of the Office communication was sent electronically on above-indicated "Notification Date" to the following e-mail address(es):

KTSDocketing2@kilpatrick.foundationip.com
ipefiling@kilpatricktownsend.com

Decision Granting Request for Prioritized Examination (Track I)	**Application No.** 16/995,607	**Applicant(s)** Lo et al.	
	Examiner IRVIN DINGLE	**Art Unit** OPET	**AIA (FITF) Status** Yes

1. THE REQUEST FILED 17 August 2020 IS **GRANTED** .

The above-identified application has met the requirements for prioritized examination

A. ☑ for an original nonprovisional application (Track I).
B. ☐ for an application undergoing continued examination (RCE).

2. **The above-identified application will undergo prioritized examination**. The application will be accorded special status throughout its entire course of prosecution until one of the following occurs:

A. filing a **petition for extension of time** to extend the time period for filing a reply;

B. filing an **amendment to amend the application to contain more than four independent claims, more than thirty total claims** , or a multiple dependent claim;

C. filing a **request for continued examination** ;

D. filing a notice of appeal;

E. filing a request for suspension of action;

F. mailing of a notice of allowance;

G. mailing of a final Office action;

H. completion of examination as defined in 37 CFR 41.102; or

I. abandonment of the application.

Telephone inquiries with regard to this decision should be directed to IRVIN DINGLE at (571)272-3210. In his/her absence, calls may be directed to Petition Help Desk at (571) 272-3282.

/IRVIN DINGLE/ Lead Paralegal Specialist, OPET	

UNITED STATES PATENT AND TRADEMARK OFFICE

UNITED STATES DEPARTMENT OF COMMERCE
United States Patent and Trademark Office
Address: COMMISSIONER FOR PATENTS
P.O. Box 1450
Alexandria, Virginia 22313-1450
www.uspto.gov

APPLICATION NUMBER	FILING or 371(c) DATE	GRP ART UNIT	FIL FEE REC'D	ATTY.DOCKET.NO	TOT CLAIMS	IND CLAIMS
16/995,607	08/17/2020		1585	080015-028410US-1199353	19	1

CONFIRMATION NO. 1256

UPDATED FILING RECEIPT

20350
KILPATRICK TOWNSEND & STOCKTON LLP
Mailstop: IP Docketing - 22
1100 Peachtree Street
Suite 2800
Atlanta, GA 30309

Date Mailed: 11/02/2020

Receipt is acknowledged of this non-provisional utility patent application. The application will be taken up for examination in due course. Applicant will be notified as to the results of the examination. Any correspondence concerning the application must include the following identification information: the U.S. APPLICATION NUMBER, FILING DATE, NAME OF FIRST INVENTOR, and TITLE OF INVENTION. Fees transmitted by check or draft are subject to collection.

Please verify the accuracy of the data presented on this receipt. If an error is noted on this Filing Receipt, please submit a written request for a corrected Filing Receipt, including a properly marked-up ADS showing the changes with strike-through for deletions and underlining for additions. If you received a "Notice to File Missing Parts" or other Notice requiring a response for this application, please submit any request for correction to this Filing Receipt with your reply to the Notice. When the USPTO processes the reply to the Notice, the USPTO will generate another Filing Receipt incorporating the requested corrections provided that the request is grantable.

Inventor(s)

Yuk-Ming Dennis Lo, Homantin, Kowloon, Hong Kong SAR, CHINA;
Rossa Wai Kwun Chiu, Shatin, New Territories, Hong Kong SAR, CHINA;
Kwan Chee Chan, Jordan, Hong Kong SAR, CHINA;
Peiyong Jiang, Shatin, New Territories, Hong Kong SAR, CHINA;
Suk Hang Cheng, Fanling, New Territories, Hong Kong SAR, CHINA;
Wenlei Peng, Shatin, New Territories, Hong Kong SAR, CHINA;
On Yee Tse, Fanling, New Territories, Hong Kong SAR, CHINA;

Applicant(s)

The Chinese University of Hong Kong, Shatin, HONG KONG;

Power of Attorney: The patent practitioners associated with Customer Number 20350

Domestic Priority data as claimed by applicant

This appln claims benefit of 63/051,210 07/13/2020
and claims benefit of 63/019,790 05/04/2020
and claims benefit of 62/991,891 03/19/2020
and claims benefit of 62/970,586 02/05/2020
and claims benefit of 62/887,987 08/16/2019

Foreign Applications for which priority is claimed (You may be eligible to benefit from the **Patent Prosecution Highway** program at the USPTO. Please see http://www.uspto.gov for more information.) - None.
Foreign application information must be provided in an Application Data Sheet in order to constitute a claim to foreign priority. See 37 CFR 1.55 and 1.76.

Permission to Access Application via Priority Document Exchange: Yes

Permission to Access Search Results: Yes

Applicant may provide or rescind an authorization for access using Form PTO/SB/39 or Form PTO/SB/69 as appropriate.

If Required, Foreign Filing License Granted: 09/24/2020

The country code and number of your priority application, to be used for filing abroad under the Paris Convention, is **US 16/995,607**

Projected Publication Date: 02/18/2021

Non-Publication Request: No

Early Publication Request: No

**** SMALL ENTITY ****

Title

DETERMINATION OF BASE MODIFICATIONS OF NUCLEIC ACIDS

Preliminary Class

Statement under 37 CFR 1.55 or 1.78 for AIA (First Inventor to File) Transition Applications: No

PROTECTING YOUR INVENTION OUTSIDE THE UNITED STATES

Since the rights granted by a U.S. patent extend only throughout the territory of the United States and have no effect in a foreign country, an inventor who wishes patent protection in another country must apply for a patent in a specific country or in regional patent offices. Applicants may wish to consider the filing of an international application under the Patent Cooperation Treaty (PCT). An international (PCT) application generally has the same effect as a regular national patent application in each PCT-member country. The PCT process **simplifies** the filing of patent applications on the same invention in member countries, but **does not result** in a grant of "an international patent" and does not eliminate the need of applicants to file additional documents and fees in countries where patent protection is desired.

Almost every country has its own patent law, and a person desiring a patent in a particular country must make an application for patent in that country in accordance with its particular laws. Since the laws of many countries differ in various respects from the patent law of the United States, applicants are advised to seek guidance from specific foreign countries to ensure that patent rights are not lost prematurely.

Applicants also are advised that in the case of inventions made in the United States, the Director of the USPTO must issue a license before applicants can apply for a patent in a foreign country. The filing of a U.S. patent application serves as a request for a foreign filing license. The application's filing receipt contains further information and guidance as to the status of applicant's license for foreign filing.

Applicants may wish to consult the USPTO booklet, "General Information Concerning Patents" (specifically, the section entitled "Treaties and Foreign Patents") for more information on timeframes and deadlines for filing foreign patent applications. The guide is available either by contacting the USPTO Contact Center at 800-786-9199, or it can be viewed on the USPTO website at http://www.uspto.gov/web/offices/pac/doc/general/index.html.

For information on preventing theft of your intellectual property (patents, trademarks and copyrights), you may wish to consult the U.S. Government website, http://www.stopfakes.gov. Part of a Department of Commerce initiative, this website includes self-help "toolkits" giving innovators guidance on how to protect intellectual property in specific countries such as China, Korea and Mexico. For questions regarding patent enforcement issues, applicants may call the U.S. Government hotline at 1-866-999-HALT (1-866-999-4258).

LICENSE FOR FOREIGN FILING UNDER

Title 35, United States Code, Section 184

Title 37, Code of Federal Regulations, 5.11 & 5.15

GRANTED

The applicant has been granted a license under 35 U.S.C. 184, if the phrase "IF REQUIRED, FOREIGN FILING LICENSE GRANTED" followed by a date appears on this form. Such licenses are issued in all applications where the conditions for issuance of a license have been met, regardless of whether or not a license may be required as set forth in 37 CFR 5.15. The scope and limitations of this license are set forth in 37 CFR 5.15(a) unless an earlier license has been issued under 37 CFR 5.15(b). The license is subject to revocation upon written notification. The date indicated is the effective date of the license, unless an earlier license of similar scope has been granted under 37 CFR 5.13 or 5.14.

This license is to be retained by the licensee and may be used at any time on or after the effective date thereof unless it is revoked. This license is automatically transferred to any related applications(s) filed under 37 CFR 1.53(d). This license is not retroactive.

The grant of a license does not in any way lessen the responsibility of a licensee for the security of the subject matter as imposed by any Government contract or the provisions of existing laws relating to espionage and the national security or the export of technical data. Licensees should apprise themselves of current regulations especially with respect to certain countries, of other agencies, particularly the Office of Defense Trade Controls, Department of State (with respect to Arms, Munitions and Implements of War (22 CFR 121-128)); the Bureau of Industry and Security, Department of Commerce (15 CFR parts 730-774); the Office of Foreign AssetsControl, Department of Treasury (31 CFR Parts 500+) and the Department of Energy.

NOT GRANTED

No license under 35 U.S.C. 184 has been granted at this time, if the phrase "IF REQUIRED, FOREIGN FILING LICENSE GRANTED" DOES NOT appear on this form. Applicant may still petition for a license under 37 CFR 5.12, if a license is desired before the expiration of 6 months from the filing date of the application. If 6 months has lapsed from the filing date of this application and the licensee has not received any indication of a secrecy order under 35 U.S.C. 181, the licensee may foreign file the application pursuant to 37 CFR 5.15(b).

SelectUSA

The United States represents the largest, most dynamic marketplace in the world and is an unparalleled location for business investment, innovation, and commercialization of new technologies. The U.S. offers tremendous resources and advantages for those who invest and manufacture goods here. Through SelectUSA, our nation works to promote and facilitate business investment. SelectUSA provides information assistance to the international investor

community; serves as an ombudsman for existing and potential investors; advocates on behalf of U.S. cities, states, and regions competing for global investment; and counsels U.S. economic development organizations on investment attraction best practices. To learn more about why the United States is the best country in the world to develop technology, manufacture products, deliver services, and grow your business, visit http://www.SelectUSA.gov or call +1-202-482-6800.

PATENT APPLICATION FEE DETERMINATION RECORD Substitute for Form PTO-875	Application or Docket Number 16/995,607

APPLICATION AS FILED - PART I

	(Column 1)	(Column 2)	SMALL ENTITY		OR	OTHER THAN SMALL ENTITY	
FOR	NUMBER FILED	NUMBER EXTRA	RATE($)	FEE($)		RATE($)	FEE($)
BASIC FEE (37 CFR 1.16(a), (b), or (c))	N/A	N/A	N/A	80		N/A	
SEARCH FEE (37 CFR 1.16(k), (i), or (m))	N/A	N/A	N/A	350		N/A	
EXAMINATION FEE (37 CFR 1.16(o), (p), or (q))	N/A	N/A	N/A	400		N/A	
TOTAL CLAIMS (37 CFR 1.16(i))	19 minus 20 =	*	x 50 =	0.00	OR		
INDEPENDENT CLAIMS (37 CFR 1.16(h))	1 minus 3 =	*	x 240 =	0.00			
APPLICATION SIZE FEE (37 CFR 1.16(s))	If the specification and drawings exceed 100 sheets of paper, the application size fee due is $310 ($155 for small entity) for each additional 50 sheets or fraction thereof. See 35 U.S.C. 41(a)(1)(G) and 37 CFR 1.16(s).			840			
MULTIPLE DEPENDENT CLAIM PRESENT (37 CFR 1.16(j))				0.00			
* If the difference in column 1 is less than zero, enter "0" in column 2.			TOTAL	1670		TOTAL	

APPLICATION AS AMENDED - PART II

		(Column 1)		(Column 2)	(Column 3)	SMALL ENTITY		OR	OTHER THAN SMALL ENTITY	
		CLAIMS REMAINING AFTER AMENDMENT		HIGHEST NUMBER PREVIOUSLY PAID FOR	PRESENT EXTRA	RATE($)	ADDITIONAL FEE($)		RATE($)	ADDITIONAL FEE($)
AMENDMENT A	Total (37 CFR 1.16(i))	*	Minus	**	=	x =		OR	x =	
	Independent (37 CFR 1.16(h))	*	Minus	***	=	x =		OR	x =	
	Application Size Fee (37 CFR 1.16(s))									
	FIRST PRESENTATION OF MULTIPLE DEPENDENT CLAIM (37 CFR 1.16(j))							OR		
						TOTAL ADD'L FEE		OR	TOTAL ADD'L FEE	

		(Column 1)		(Column 2)	(Column 3)					
		CLAIMS REMAINING AFTER AMENDMENT		HIGHEST NUMBER PREVIOUSLY PAID FOR	PRESENT EXTRA	RATE($)	ADDITIONAL FEE($)		RATE($)	ADDITIONAL FEE($)
AMENDMENT B	Total (37 CFR 1.16(i))	*	Minus	**	=	x =		OR	x =	
	Independent (37 CFR 1.16(h))	*	Minus	***	=	x =		OR	x =	
	Application Size Fee (37 CFR 1.16(s))									
	FIRST PRESENTATION OF MULTIPLE DEPENDENT CLAIM (37 CFR 1.16(j))							OR		
						TOTAL ADD'L FEE		OR	TOTAL ADD'L FEE	

* If the entry in column 1 is less than the entry in column 2, write "0" in column 3.
** If the "Highest Number Previously Paid For" IN THIS SPACE is less than 20, enter "20".
*** If the "Highest Number Previously Paid For" IN THIS SPACE is less than 3, enter "3".
The "Highest Number Previously Paid For" (Total or Independent) is the highest found in the appropriate box in column 1.

==========
Sequence Listing was accepted.
If you need help call the Patent Electronic Business Center at (866) 217-9197 (toll free).
Reviewer: Zhang, Yizhu (ASRC)
Timestamp: [year=2020; month=10; day=29; hr=15; min=32; sec=12; ms=915;]
==========

Validated By CRFValidator v 1.0.5

Application No: 16995607 Version No: 1.0

Input Set:

Output Set:

Started: 2020-10-27 16:44:16.784
Finished: 2020-10-27 16:44:16.909
Elapsed: 0 hr(s) 0 min(s) 0 sec(s) 125 ms
Total Warnings: 5
Total Errors: 0
No. of SeqIDs Defined: 5
Actual SeqID Count: 5

Error code	Error Description
W 213	Artificial or Unknown found in <213> in SEQ ID (1)
W 213	Artificial or Unknown found in <213> in SEQ ID (2)
W 213	Artificial or Unknown found in <213> in SEQ ID (3)
W 213	Artificial or Unknown found in <213> in SEQ ID (4)
W 213	Artificial or Unknown found in <213> in SEQ ID (5)

SEQUENCE LISTING

<110> THE CHINESE UNIVERSITY OF HONG KONG

<120> DETERMINATION OF BASE MODIFICATIONS OF NUCLEIC ACIDS

<130> 080015-028410US-1199353

<140> US 16/995,607
<141> 2020-08-17

<150> 63/051,210
<151> 2020-07-13

<150> 63/019,790
<151> 2020-05-04

<150> 62/991,891
<151> 2020-03-19

<150> 62/970,586
<151> 2020-02-05

<150> 62/887,987
<151> 2019-08-16

<160> 5

<170> PatentIn version 3.5

<210> 1
<211> 36
<212> RNA
<213> Artificial Sequence

<220>
<223> Description of Artificial Sequence: Synthetic
oligonucleotide

<400> 1
gccuguaauc ccagcacuuu guuuuagagc uaugcu 36

<210> 2
<211> 67
<212> RNA
<213> Artificial Sequence

<220>
<223> Description of Artificial Sequence: Synthetic
oligonucleotide

<400> 2
agcauagcaa guuaaaauaa ggcuaguccg uuaucaacuu gaaaaagugg caccgagucg 60

gugcuuu 67

<210> 3
<211> 36
<212> RNA
<213> Artificial Sequence

<220>
<223> Description of Artificial Sequence: Synthetic
oligonucleotide

<400> 3
agggucucgc ucugucgccc guuuuagagc uaugcu 36

<210> 4
<211> 10
<212> DNA
<213> Artificial Sequence

<220>
<223> Description of Artificial Sequence: Synthetic
oligonucleotide

<400> 4
atacgtacgt 10

<210> 5
<211> 10
<212> DNA
<213> Artificial Sequence

<220>
<223> Description of Artificial Sequence: Synthetic
oligonucleotide

<400> 5
atacgtacgt 10

INFORMATION DISCLOSURE STATEMENT BY APPLICANT (Not for submission under 37 CFR 1.99)		
	Application Number	16/995,607
	Filing Date	August 17, 2020
	First Named Inventor	Yuk-Ming Dennis Lo
	Art Unit	Not Yet Known
	Examiner Name	Not Yet Known
	Attorney Docket Number	080015-028410US-1199353

U.S. PATENTS

Examiner Initial*	Cite No	Patent Number	Kind Code [1]	Issue Date	Name of Patentee or Applicant of cited Document	Pages, Columns, Lines, Where Relevant Passages or Relevant Figures Appear
	1.	9,116,118	B2	08-25-2015	Turner et al.	
	2.	9,175,338	B2	11-03-2015	Flusberg et al.	
	3.	10,590,484	B2	03-17-2020	Korlach et al.	
	4.	8,467,976	B2	06-18-2013	Lo et al.	
	5.	10,468,121	B2	06-18-2013	Lo et al.	
	6.	10,465,232	B1	11-05-2019	Wu et al.	
	7.	10,438,691	B2	10-08-2019	Kim et al.	
	8.	9,175,341	B2	11-03-2015	Flusberg et al.	

U.S. PATENT APPLICATION PUBLICATIONS

Examiner Initial*	Cite No	Publication Number	Kind Code [1]	Publication Date	Name of Patentee or Applicant of cited Document	Pages, Columns, Lines, Where Relevant Passages or Relevant Figures Appear
	9.	2011/0183320	A1	07-28-2011	Flusberg et al.	
	10.	2017/0037464	A1	02-09-2017	Turner et al.	
	11.	2017/0233802	A1	08-17-2017	Flusberg et al.	
	12.	2018/0201993	A1	07-19-2018	Turner et al.	
	13.	2019/0024162	A1	01-24-2019	Korlach et al.	
	14.	2017/0159132	A1	06-08-2017	Okino et al.	
	15.	2014/0004511	A1	01-02-2014	Korlach et al.	

FOREIGN PATENT DOCUMENTS

Examiner Initial*	Cite No	Foreign Document Number [3]	Country Code [2]	Kind Code [4]	Publication Date	Name of Patentee or Applicant of cited Document	Pages, Columns, Lines where Relevant Passages or Relevant Figures Appear	T [5]
	16.	101896167	KR	B1	09-07-2018		English abstract attached. Equivalent to WO 2009013492 A1 below.	
	17.	2009013492	WO	A1	01-29-2009	The Chinese University of Hong Kong	Equivalent to KR 101896167 B1 above.	

NON-PATENT LITERATURE DOCUMENTS

Examiner Initials*	Cite No	Include name of the author (in CAPITAL LETTERS), title of the article (when appropriate), title of the item (book, magazine, journal, serial, symposium, catalog, etc.), date, page(s), volume-issue number(s), publisher, city and/or country where published.	T [5]

INFORMATION DISCLOSURE STATEMENT BY APPLICANT (Not for submission under 37 CFR 1.99)		
	Application Number	16/995,607
	Filing Date	August 17, 2020
	First Named Inventor	Yuk-Ming Dennis Lo
	Art Unit	Not Yet Known
	Examiner Name	Not Yet Known
	Attorney Docket Number	080015-028410US-1199353

	18.	FLUSBERG, Benjamin A. et al.; "Direct detection of DNA methylation during single-molecule, real-time sequencing"; Nature Methods; June 2010; Vol. 7, No. 6; pages 461-465 (7 pages).	
	19.	LIU, Qian et al.; "Detection of DNA base modifications by deep recurrent neural network on Oxford Nanopore sequencing data"; Nature Communications; 2019; 10(1):2449; doi: 10.1038/s41467-019-10168-2; 11 pages.	
	20.	NI, Peng et al.; "DeepSignal: detecting DNA methylation state from Nanopore sequencing reads using deep-learning"; Bioinformatics; 2019; Vol. 35, No. 22; doi: 10.1093/bioinformatics/btz276; 10 pages.	
	21.	SUZUKI, Yuta et al.; "AgIn: measuring the landscape of CpG methylation of individual repetitive elements"; Bioinformatics; 2016; Vol. 32, No. 19; pages 2911-2919.	
	22.	BLOW, Matthew J. et al.; "The Epigenomic Landscape of Prokaryotes"; PLOS Genetics; February 12, 2016; 12(2): e1005854; doi:10.1371/journal.pgen.1005854; 28 pages.	
	23.	CLARK, Tyson A. et al.; "Enhanced 5-methylcytosine detection in single-molecule, real-time sequencing via Tet1 oxidation"; BMC Biology; 2013; Vol. 11, No. 4; http://www.biomedcentral.com/1741-7007/11/4; 10 pages.	
	24.	CLARK, Tyson A. et al.; "Characterization of DNA methyltransferase specificities using single-molecule, real-time DNA sequencing"; Nucleic Acids Research; 2012; Vol. 40, No. 4; e29; https://doi.org/10.1093/nar/gkr1146; 12 pages.	
	25.	EID, John et al.; "Real-Time DNA Sequencing from Single Polymerase Molecules"; Science; 02 January 2009; Volume 323, Issue 5910; DOI: 10.1126/science.1162986; page 133-138 (7 pages).	
	26.	FENG, Zhixing et al.; "Detecting DNA Modifications from SMRT Sequencing Data by Modeling Sequence Context Dependence of Polymerase Kinetic"; PLOS Computational Biology; March 2013; Volume 9, Issue 3; e1002935; https://doi.org/10.1371/journal.pcbi.1002935; 10 pages.	
	27.	LIU, Yibin et al.; "Bisulfite-free direct detection of 5-methylcytosine and 5-hydroxymethylcytosine at base resolution"; Nature Biotechnology Letters; April 2019; Volume 37; pages 424-429 (11 pages).	
	28.	SCHADT, Eric E. et al.; "Modeling kinetic rate variation in third generation DNA sequencing data to detect putative modifications to DNA bases"; Genome Research; 2013; Vol. 23, No. 1; doi: 10.1101/gr.136739.111; pages 129-141.	
	29.	FENG, Zhixing et al.; "qDNAmod: a statistical model-based tool to reveal intercellular heterogeneity of DNA modification from SMRT sequencing data"; Nucleic Acids Research; 2014; Vol. 42, No. 22; doi: 10.1093/nar/gku1097; pages 13488-13499.	

INFORMATION DISCLOSURE STATEMENT BY APPLICANT (Not for submission under 37 CFR 1.99)		
	Application Number	16/995,607
	Filing Date	August 17, 2020
	First Named Inventor	Yuk-Ming Dennis Lo
	Art Unit	Not Yet Known
	Examiner Name	Not Yet Known
	Attorney Docket Number	080015-028410US-1199353

	30.	SUZUKI, Yuta; "Observing comprehensive DNA methylomes via single molecule real-time sequencing: application to diploid and centromeric methylation"; dissertation submitted to The Graduate School of Frontier Sciences, The University of Tokyo, on December 13, 2017; 118 pages.	

EXAMINER SIGNATURE			
Examiner Signature		Date Considered	

*EXAMINER: Initial if reference considered, whether or not citation is in conformance with MPEP 609. Draw line through a citation if not in conformance and not considered. Include copy of this form with next communication to applicant.

[1] See Kind Codes of USPTO Patent Documents at www.USPTO.GOV or MPEP 901.04. [2] Enter office that issued the document, by the two-letter code (WIPO Standard ST.3). [3] For Japanese patent documents, the indication of the year of the reign of the Emperor must precede the serial number of the patent document.
[4] Kind of document by the appropriate symbols as indicated on the document under WIPO Standard ST.16 if possible. [5] Applicant is to place a check mark here if English language translation is attached.

I hereby certify that this correspondence is being filed via EFS-Web with the United States Patent and Trademark Office on October 27, 2020.

KILPATRICK TOWNSEND & STOCKTON LLP

By: /Heather A. Orosco/
Heather A. Orosco

PATENT
Attorney Docket No.: 080015-028410US-1199353
Client Ref. No.: 19/MED/901 US Reg

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

In re Application of:	Confirmation No.: 1256
Yuk-Ming Dennis Lo	Examiner: Not Yet Known
Application No.: 16/995,607	Technology Center/Art Unit: Not Yet Known
Filed: August 17, 2020	
For: DETERMINATION OF BASE MODIFICATIONS OF NUCLEIC ACIDS	INFORMATION DISCLOSURE STATEMENT
Customer No.: 20350	

Commissioner for Patents
P.O. Box 1450
Alexandria, VA 22313-1450

Commissioner:

The references cited on attached form PTO/SB/08A are being called to the attention of the Examiner. In accordance with the provisions of 37 CFR §1.98(a)(2), copies of any cited U.S. Patent and U.S. Patent Application Publications are not provided. Copies of the remaining cited references are provided.

It is respectfully requested that the cited references be expressly considered during the prosecution of this application, and the references be made of record therein and appear among the "references cited" on any patent to issue therefrom.

As provided for by 37 CFR §1.97(g) and (h), no inference should be made that the information and references cited are prior art merely because they are in this statement and no

representation is being made that a search has been conducted or that this statement encompasses all the possible relevant information.

This IDS is being filed before the mailing date of the first Office Action on the merits.

Applicant believes that no fee is required for submission of this statement. However, if any additional fees are due for the submission of this Information Disclosure Statement, please deduct those fees from Deposit Account No. 20-1430.

Respectfully submitted,

/David D. Hsu/

David D. Hsu
Registration No. 71659

KILPATRICK TOWNSEND & STOCKTON LLP

Attachment

73961535V.1

Espacenet

Bibliographic data: KR101896167 (B1) — 2018-09-07

DETERMINING A NUCLEIC ACID SEQUENCE IMBALANCE

Inventor(s): 로, 육-밍 데니스, ; 치우, 로사 와이 큔, ; 찬, 콴 치, ; 지, 베니 충 잉, ; 총, 카 춘

Applicant(s): 더 차이니즈 유니버시티 오브 홍콩

Classification:
- **international:** *C12Q1/68; G06F19/18; G06F19/22*
- **cooperative:** **default** C12Q1/68 (EA); C12Q1/6809 (CN); C12Q1/6827 (EP, CN, KR, US); C12Q1/6869 (CN); C12Q1/6883 (EP, CN, KR, US); C12Q1/6888 (US); G01N33/48 (EA); G01N33/50 (EA); G16B15/00 (KR); G16B20/00 (CN, KR, US); G16B20/10 (EP); G16B30/00 (CN, KR); C12Q2600/112 (KR, US); C12Q2600/154 (EP, CN, KR, US); C12Q2600/156 (KR, US); G01N2800/387 (KR, US); G16B30/00 (EP, US); Y02A90/10 (EP, US)

C-sets
- **C12Q1/6827, C12Q2521/331, C12Q2537/16 (EP, CN, US);**
- **C12Q1/6827, C12Q2537/16 (CN);**
- **C12Q1/6869, C12Q2535/122 (CN)**

less

Application number: KR20167021213 20080723 Global Dossier

Priority number(s): US20070951438P 20070723 ; WO2008GB02524 20080723

Also published as: AU2008278839 (A1) AU2008278839 (B2) AU2008278843 (A1) BRPI0814670 (A2) BRPI0814670 (B1) CA2693081 (A1) CA2693081 (C) CA2694007 (A1) CA2694007 (C) CA2900927 (A1) CA2900927 (C) CA3009992 (A1) CA3029497 (A1) CA3076142 (A1) CA3076159 (A1) CN101849236 (A) CN101971178 (A) CN101971178 (B) CN103849684 (A) CN103853916 (A) CN103853916 (B) CN103902809 (A) CN103902809 (B) CN106676188 (A) CN106834474 (A) CN106834474 (B) CN106834481 (A) CN106886688 (A) CN106886688 (B) CN107083424 (A) CN107083425 (A) CY1114773 (T1) CY1117525 (T1) DK2183693 (T3) DK2183693 (T4) DK2183693 (T5) DK2514842 (T3) DK2527471 (T3) DK2557519 (T3) EA017966 (B1) EA028642 (B1) EA035451 (B1) EA035451 (B9) EA201000231 (A1) EA201201551 (A1) EA201300072 (A1) EA201600280 (A1) EA201791612 (A2) EA201791612 (A3) EP2183692 (A1) EP2183692 (B1) EP2183693 (A1) EP2183693 (B1) EP2183693 (B2) EP2514842 (A2) EP2514842 (A3) EP2514842 (B1) EP2527471 (A2) EP2527471 (A3) EP2527471 (B1) EP2557517 (A2) EP2557517 (A3) EP2557518 (A2) EP2557518 (A3) EP2557518 (B1) EP2557519 (A2) EP2557519 (A3) EP2557519 (B1) EP2557520 (A2) EP2557520 (A3) EP3067807 (A1) EP3540739 (A1) EP3656870 (A1) ES2441807 (T3) ES2441807 (T5) ES2571738 (T3) HK1144024 (A1) HK1177768 (A1) HK1182195 (A1) HK1199067 (A1) HK1224033 (A1) HRP20140009 (T1) HRP20140009 (T4) HRP20160493 (T1) HUE030510 (T2) IL203311 (A) IL233261 (A) JP2010534068 (A) JP2010534069 (A) JP2014073134 (A) JP2015142588 (A) JP2016185162 (A) JP2017148073 (A) JP2019000113 (A) JP2019013245 (A) JP2020031663 (A) JP2020031664 (A) JP2020115887 (A) JP5519500 (B2) JP5736170 (B2) JP6151739 (B2) JP6383837 (B2) JP6522554 (B2) JP6629940 (B2) JP6695392 (B2) KR101646978 (B1) KR101829564 (B1) KR101829565 (B1) KR101916456 (B1) KR101966262 (B1) KR101972994 (B1) KR102060911 (B1) KR102076438 (B1) KR102112438 (B1) KR102128960 (B1) KR102147626 (B1) KR20100058503 (A) KR20100075826 (A) KR20160030404 (A) KR20160113145 (A) KR20160113146 (A) KR20160113147 (A) KR20160113148 (A) KR20170127073 (A) KR20180100713 (A) KR20180114251 (A) KR20180121695 (A) KR20190114039 (A) KR20190114041 (A) KR20190143494 (A) KR20200055151 (A) KR20200100860 (A) MX2010000846 (A) MX341573 (B) MX346069 (B) NZ582702 (A) NZ600407 (A) PL2183693 (T3) PL2183693 (T4) PL2183693 (T5) PL2514842 (T3) PT2183693 (E) SG183062 (A1) SI2183693 (T1) SI2183693 (T2) SI2514842 (T1) US10208348 (B2) US2009029377 (A1) US2009087847 (A1) US2014045181 (A1) US2014256559 (A1) US2014256560 (A1) US2014329695 (A1) US2014329696 (A1) US2019136323 (A1) US2020056242 (A1) US8706422 (B2) US8972202 (B2) US9051616 (B2) WO2009013492 (A1) WO2009013496 (A1) ZA201000524 (B) less

Abstract not available for KR101896167 (B1)
Abstract of corresponding document: WO2009013492 (A1)

Methods, systems, and apparatus are provided for determining whether a nucleic acid sequence imbalance exists within a biological sample. One or more cutoff values for determining an imbalance of, for example, the ratio of the two sequences (or sets of sequences) are chosen. The cutoff value may be determined based at least in part on the percentage of fetal DNA in a sample, such as maternal plasma, containing a background of maternal nucleic acid sequences. The cutoff value may also be determined based on an average concentration of a sequence per reaction. In one aspect, the cutoff value is determined from a proportion of informative wells that are estimated to contain a particular nucleic acid sequence, where the proportion is determined based on the above-mentioned percentage and/or average concentration. The cutoff value may be determined using many different types of methods, such as sequential probability ratio testing (SPRT).

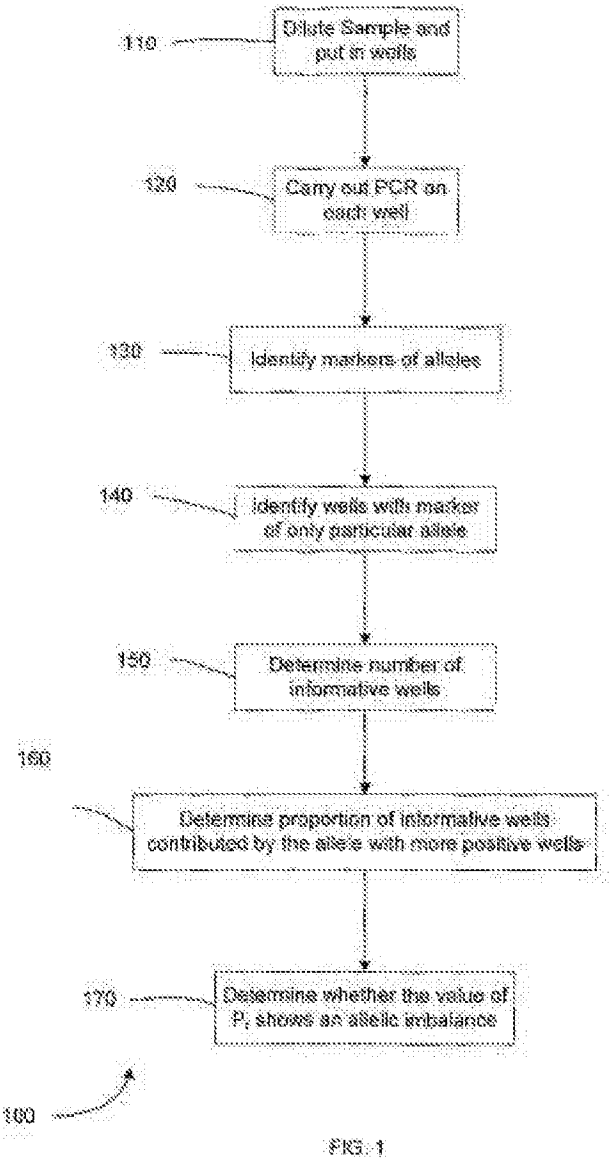


FIG. 1

(19) 대한민국특허청(KR)
(12) 등록특허공보(B1)

(45) 공고일자 2018년09월07일
(11) 등록번호 10-1896167
(24) 등록일자 2018년08월31일

(51) 국제특허분류(Int. Cl.)
G06F 19/18 (2011.01) *C12Q 1/68* (2018.01)
G06F 19/22 (2017.01)
(52) CPC특허분류
G06F 19/18 (2013.01)
C12Q 1/6827 (2018.05)
(21) 출원번호 10-2016-7021213(분할)
(22) 출원일자(국제) 2008년07월23일
심사청구일자 2016년09월01일
(85) 번역문제출일자 2016년08월02일
(65) 공개번호 10-2016-0113147
(43) 공개일자 2016년09월28일
(62) 원출원 특허 10-2010-7003906
원출원일자(국제) 2008년07월23일
심사청구일자 2013년07월23일
(86) 국제출원번호 PCT/GB2008/002524
(87) 국제공개번호 WO 2009/013492
국제공개일자 2009년01월29일
(30) 우선권주장
60/951,438 2007년07월23일 미국(US)
(56) 선행기술조사문헌
JP2007515947 A*
KR1020020064298 A*
WO2007075836 A2*
*는 심사관에 의하여 인용된 문헌

(73) 특허권자
더 차이니즈 유니버시티 오브 홍콩
중국 홍콩 엔. 티. 새딘 더 차이니즈 유니버시티 오브 홍콩
(72) 발명자
로, 육-밍 데니스
홍콩, 구룡, 호만틴, 7 킹 탁 스트리트, 4 플로어
치우, 로사 와이 쿤
홍콩, 뉴 테리토리스, 타이 포, 1 홍 람 드라이브, 콘스텔레이션 코브, 블록 1, 플랫 1에이
(뒷면에 계속)
(74) 대리인
손민

전체 청구항 수 : 총 20 항 심사관 : 박진아

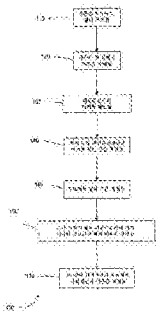
(54) 발명의 명칭 **핵산 서열 불균형의 결정**

(57) 요 약

본 발명에서는, 핵산 서열 불균형이 생물학적 샘플 내에 존재하는지 여부를 결정하는 방법, 시스템 및 장치가 제공된다. 예를 들어, 2개의 서열 (또는 서열 세트)의 비의 불균형을 측정하기 위한 하나 이상의 분별값이 선택된다. 분별값은 적어도 부분적으로 샘플, 예컨대 모체 핵산 서열의 배경을 함유하는 모체 혈장 중 태아 DNA의 백

(뒷면에 계속)

대 표 도 - 도1



분율을 기준으로 결정될 수 있다. 분별값은 또한 반응 당 서열의 평균 농도를 기준으로 결정될 수 있다. 한 측면에서, 분별값은 특정 핵산 서열을 함유하는 것으로 평가되는 정보제공 웰의 비율로부터 결정되며, 그 비율은 상기 언급된 백분율 및/또는 평균 농도를 기준으로 결정된다. 분별값은 다수의 상이한 유형의 방법, 예컨대 순차 확률비 검증 (SPRT)을 사용하여 결정될 수 있다.

(52) CPC특허분류
C12Q 1/6883 (2018.05)
C12Q 1/6888 (2018.05)
G06F 19/22 (2018.05)
C12Q 2600/112 (2013.01)
C12Q 2600/154 (2013.01)
C12Q 2600/156 (2013.01)
G01N 2800/387 (2013.01)

(72) 발명자
찬, 콴 치
홍콩, 구룡, 메이 푸 신 추엔, 브로드웨이 스트리트, 블록 34, 13/에프, 플랫 에이
지, 베니 충 잉
홍콩, 뉴 테리토리스, 마 온 산, 라 코스타, 타워 2, 플렛 18이
총, 카 춘
홍콩, 뉴 테리토리스, 타이 포, 푸 신 에스테이트, 신 킹 하우스, 29/에프, 플랫 06

명 세 서

청구범위

청구항 1

핵산 서열 불균형이 생물학적 샘플 내에 존재하는지 여부를 결정하기 위한 작업을 수행하기 위한 컴퓨터 시스템 제어용의 다수의 지시로 코딩된 비일시적 컴퓨터 판독가능 기록 매체로서, 상기 지시는 하기를 포함하는 것인, 비일시적 컴퓨터 판독가능 기록 매체:

다수의 반응으로부터 데이터를 수득하는 단계로, 상기 데이터는

(1) 생물학적 샘플 중의 임상적으로 관련된 핵산 서열의 제1 양을 나타내는 제1 세트의 정량 데이터; 및

(2) 임상적으로 관련된 핵산 서열과 상이한 생물학적 샘플 중의 배경 핵산 서열의 제2 양을 나타내는 제2 세트의 정량 데이터를 포함하고,

상기 임상적으로 관련된 핵산 서열과 배경 핵산 서열이 제1 유형의 세포 및 하나 이상의 제2 유형의 세포로부터 유래된 것인, 단계;

두 데이터 세트로부터 파라미터를 결정하는 단계;

생물학적 샘플에서 제1 유형의 세포로부터의 핵산 서열의 양을 측정하여 수득된 제1 백분율로부터 제1 분별값(cutoff value)을 유도하는 단계;

파라미터를 제1 분별값과 비교하는 단계; 및

비교를 기준으로, 생물학적 샘플 내에 핵산 서열 불균형이 존재하는지 여부의 분류를 결정하는 단계.

청구항 2

제1항에 있어서, 서열 불균형이 종양에서 염색체의 획득 또는 소실과 관련되는 것인, 비일시적 컴퓨터 판독가능 기록 매체.

청구항 3

제1항에 있어서, 제1 유형의 세포가 제1 유기체로부터의 것이고, 제2 유형의 세포가 제2 유기체로부터의 것인, 비일시적 컴퓨터 판독가능 기록 매체.

청구항 4

제3항에 있어서, 제1 유기체가 제2 유기체인 태아를 임신한 여성 대상체이고, 제1 백분율의 결정이 여성 대상체와 태아 사이의 다형성 차이의 정량화를 포함하는 것인, 비일시적 컴퓨터 판독가능 기록 매체.

청구항 5

제3항에 있어서, 제1 유기체가 제2 유기체인 태아를 임신한 여성 대상체이고, 태아가 남성이며, 생물학적 샘플 중의 태아 DNA의 분획 농도를 결정하는 단계가 Y 염색체 DNA 농도를 결정하는 단계를 포함하는 것인, 비일시적 컴퓨터 판독가능 기록 매체.

청구항 6

제3항에 있어서, 제1 유기체가 제2 유기체인 태아를 임신한 여성 대상체이고, 생물학적 샘플 중의 태아 DNA의 분획 농도를 결정하는 단계가 제1 유전자좌에서의 태아-특이적 메틸화 패턴을 나타내는 DNA 분자의 양을 제1 유전자좌에서의 DNA 분자의 총량과 비교하는 단계를 포함하는 것인, 비일시적 컴퓨터 판독가능 기록 매체.

청구항 7

제6항에 있어서, 태아-유도된 DNA 분자가 과-메틸화되고, 모체-유도된 DNA 분자가 저-메틸화되는 것인, 비일시

적 컴퓨터 판독가능 기록 매체.

청구항 8

제1항에 있어서, 상기 지시가

반응 당, 과소발현된 임상적으로 관련된 핵산 서열 또는 배경 핵산 서열인, 참조 핵산 서열의 제1 평균 농도를 결정하는 단계; 및

제1 평균 농도에 제1 백분율로부터 유도된 인자를 곱하여 참조 핵산 서열이 아닌 핵산 서열의 제2 평균 농도를 얻는 단계를 추가로 포함하는 것인, 비일시적 컴퓨터 판독가능 기록 매체.

청구항 9

제8항에 있어서, 상기 지시가 참조 핵산 서열에 대한 데이타로부터 유도된 값이 입력된 역 확률 분포를 사용하여 다수의 반응 각각에서 참조 핵산 서열의 평균 농도를 결정하는 단계를 더 포함하는 것인, 비일시적 컴퓨터 판독가능 기록 매체.

청구항 10

태아를 임신한 여성 대상체의 생물학적 샘플 내의 태아 DNA의 분획 농도를 결정하는 방법으로서, 생물학적 샘플은 여성 대상체 및 태아로부터 핵산 분자를 포함하고, 모체는 제1 대립유전자에 대한 제1 유전자좌에서 동형접합성이고 태아는 제1 대립유전자 및 제1 대립유전자와 상이한 제2 대립유전자에 대한 제1 유전자좌에서 이형접합성이며, 상기 방법은 하기를 포함하는 것인, 방법:

생물학적 샘플로부터의 핵산 분자를 포함하는 제1의 다수의 반응으로부터 제1 데이터를 수득하는 단계로, 상기 반응은 다수의 관심 폴리뉴클레오티드 서열의 존재 또는 부재를 나타내고, 상기 데이터는

(1) 제1 대립유전자에 대해 양성인 반응의 제1 수를 나타내는 제1 세트의 정량 데이터; 및

(2) 제2 대립유전자에 대해 양성인 반응의 제2 수를 나타내는 제2 세트의 정량 데이터를 포함하는, 단계; 및

제1 수를 제2 수와 비교하여 태아 DNA의 분획 농도를 결정하는 단계.

청구항 11

제10항에 있어서, 태아 DNA의 분획 농도가 제1 수 및 제2 수를 사용하여 백분율로서 결정되는 것인 방법.

청구항 12

제10항에 있어서, 하기 단계를 더 포함하는 것인 방법:

생물학적 샘플로부터의 핵산 분자를 포함하는 제2의 다수의 반응으로부터 제2 데이터를 수득하는 단계로, 상기 데이터는

(1) 제1 유전자좌를 포함하지 않는 하나 이상의 임상적으로 관련된 핵산 서열의 제3 양을 나타내는 제3 세트의 정량 데이터; 및

(2) 하나 이상의 임상적으로 관련된 핵산 서열과 상이한 하나 이상의 배경 핵산 서열의 제4 양을 나타내는 제4 세트의 정량 데이터를 포함하고,

상기 임상적으로 관련된 핵산 서열 및 배경 핵산 서열은 여성 대상체 및 태아로부터의 핵산 분자로부터 유래된 것인, 단계;

제3 및 제4 데이터 세트로부터 제3 양과 제4 양 사이의 상대적 양을 제공하는 파라미터를 결정하는 단계;

태아 DNA의 분획 농도로부터 제1 분별값을 유도하는 단계;

파라미터를 제1 분별값과 비교하는 단계; 및

비교를 기준으로, 핵산 서열 불균형이 존재하는지 여부의 분류를 결정하는 단계.

청구항 13

제12항에 있어서, 서열 분균형이 염색체 이수성(aneuploidy)인 방법.

청구항 14

제12항에 있어서, 임상적으로 관련된 핵산 서열 중 하나가 유전자 다형성의 대립유전자이고, 배경 핵산 서열 중 하나가 유전자 다형성의 또 다른 대립유전자인 방법.

청구항 15

제12항에 있어서, 제2의 다수의 반응을 수행하기 전에 하나 이상의 임상적으로 관련된 핵산 서열 및 하나 이상의 배경 핵산 서열에 대해 생물학적 샘플을 농축시키는 단계를 더 포함하는 방법.

청구항 16

제15항에 있어서, 하나 이상의 임상적으로 관련된 핵산 서열 및 하나 이상의 배경 핵산 서열에 대해 생물학적 샘플을 농축시키는 단계가 혼성화로 하나 이상의 임상적으로 관련된 핵산 서열 및 하나 이상의 배경 핵산 서열을 선별하기 위해 올리고뉴클레오티드 어레이를 사용하는 단계를 포함하는 것인 방법.

청구항 17

제15항에 있어서, 생물학적 샘플을 농축시키는 단계가 하나 이상의 임상적으로 관련된 핵산 서열 및 하나 이상의 배경 핵산 서열로부터 DNA를 증폭시키는 단계를 포함하는 것인 방법.

청구항 18

제12항에 있어서, 태아 DNA의 분획 농도가 파라미터 결정 후 결정되는 것인 방법.

청구항 19

제12항에 있어서, 제1의 다수의 반응이 제2의 다수의 반응과 동일한 것인 방법.

청구항 20

제10항에 있어서, 반응이 서열분석 반응 또는 증폭 반응인 방법.

발명의 설명

기 술 분 야

[0001] 우선권 주장

[0002] 본 출원은 2007년 7월 23일 출원된, 발명의 명칭이 "핵산 서열 불균형의 결정(DETERMINING A NUCLEIC ACID SEQUENCE IMBALANCE)"인 미국 가출원 번호 제60/951438호 (사건관리 번호 제016285-005200US호)를 우선권으로 주장하며 이의 정규출원이고, 상기 가출원의 전체 내용은 모든 목적을 위해 참조로 본원에 포함된다.

[0003] 관련 출원의 상호 참조

[0004] 본 출원은 또한 동시에 출원된, 발명의 명칭이 "게놈 서열분석을 사용한 태아 염색체 이수성의 진단(DIAGNOSING FETAL CHROMOSOMAL ANEUPLOIDY USING GENOMIC SEQUENCING)"인 정규출원 (사건관리 번호 제016285-005220US호)에 관한 것이고, 이 가출원의 전체 내용은 모든 목적을 위해 참조로 본원에 포함된다.

[0005] 발명의 분야

[0006] 본 발명은 일반적으로 2개의 상이한 핵산 서열간 불균형을 결정하는 유전사형 및 질환의 진단 시험, 보다 특히 모체 혈액 샘플의 시험을 통한 태아에서의 다운 증후군, 다른 염색체 이수성(aneuploidy), 돌연변이 및 유전자형의 확인에 관한 것이다. 본 발명은 또한 암의 검출, 이식의 모니터링 및 감염성 질환의 모니터링에 관한 것이다.

배경기술

[0008] 유전 질환, 암 및 다른 상태는 종종 2개의 상응하는 염색체 또는 대립유전자 또는 다른 핵산 서열에서의 불균형으로부터 초래되거나 이러한 불균형을 생성한다. 즉, 다른 서열에 비해 한 서열의 양이 정상적인 경우보다 많거나 적다. 통상적으로, 정상 비율은 대등한 50/50 비율이다. 다운 증후군 (삼염색체 21)은 추가 염색체 21의 불균형을 갖는 질환이다.

[0009] 삼염색체 21의 통상적인 출생 전 진단 방법은 한정된 태아 소실 위험을 낳는 침습적 절차, 예컨대 양수검사 또는 융모막 융모 생검에 의해 태아 물질을 샘플링하는 것을 포함한다. 비-침습적 절차, 예컨대 초음파검사 및 생화학 마커에 의한 스크리닝은 한정적인 침습적 진단 절차 전에 임신 여성을 위험-등급화하는데 사용되었다. 그러나, 이러한 스크리닝 방법은 전형적으로 핵심 염색체 이상 대신에 삼염색체 21과 관련된 부대징후를 측정하므로, 최적에 못미치는 진단 정확성 및 다른 단점들 (예컨대 재태 기간에 의해 많은 영향을 받음)을 갖는다.

[0010] 1997년에 모체 혈장 중 순환 무-세포 태아 DNA의 발견은 비-침습적 출생 전 진단의 새로운 가능성을 제공하였다 (문헌 [Lo, YMD and Chiu, RWK 2007 Nat Rev Genet 8, 71-77]). 이 방법이 성별-연계된 유전자 장애 (문헌 [Costa, JM et al. 2002 N Engl J Med 346, 1502]) 및 특정 단일 유전자 장애 (문헌 [Lo, YMD et al. 1998 N Engl J Med 339, 1734-1738])의 출생 전 진단에 용이하게 적용되었지만, 태아 염색체 이수성의 출생 전 검출에 적용하기에는 상당한 난점이 있었다 (상기 문헌 [Lo, YMD and Chiu, RWK 2007]). 첫째, 태아 핵산은 모체 혈장 중에서, 종종 분석에 방해가 될 수 있는 다량의 모체 기원 배경 핵산과 공존한다 (문헌 [Lo, YMD et al. 1998 Am J Hum Genet 62, 768-775]). 둘째, 태아 핵산은 모체 혈장 중에서 우세하게는 무-세포 형태로 순환하여, 태아 게놈 내 유전자 또는 염색체의 투여 정보 유도를 어렵게 만든다.

[0011] 이들 난점을 극복하는 유의한 개발이 최근에 이루어졌다 (문헌 [Benachi, A & Costa, JM 2007 Lancet 369, 440-442]). 한 접근법은 모체 혈장 중 태아-특이적 핵산을 검출하여, 모체 배경 방해 문제를 극복하는 것이다 (상기 문헌 [Lo, YMD and Chiu, RWK 2007]). 염색체 21의 투여는 태반-유래 DNA/RNA 분자에서 다형성 대립유전자의 비율로부터 추론되었다. 그러나, 이러한 방법은, 샘플이 보다 적은 양의 표적 유전자를 함유하는 경우 덜 정확하며, 하나의 다형성이 사용되는 경우 단지 그 집단의 아군인, 표적 다형성에 대해 이형접합성인 태아에 개만 적용될 수 있다.

[0012] 달란 등(Dhallan et al)의 문헌 [Dhallan, R, et al. 2007, supra Dhallan, R, et al. 2007 Lancet 369, 474-481]에는, 모체 혈장에 포름알데히드를 첨가함으로써 순환 태아 DNA의 비율을 증가시키는 별법의 전략이 기재되어 있다. 모체 혈장 중 태아에 의해 기여된 염색체 21 서열의 비율은, 염색체 21 상의 단일 뉴클레오티드 다형성 (SNP)에 대한 부계-유전 태아-특이적 대립유전자 대 비-태아-특이적 대립유전자의 비율을 평가함으로써 결정되었다. SNP 비율은 참조 염색체에 대해 유사하게 추정되었다. 이어서, 태아 염색체 21의 불균형은, 염색체 21에 대한 SNP 비율과 참조 염색체에 대한 SNP 비율 사이의 통계적으로 유의한 차이를 검출함으로써 추론되었으며, 상기 유의성은 0.05 이하의 고정 p-값을 사용하여 정의된다. 고 집단 적용범위를 보장하기 위해, 염색체당 500개 초과의 SNP가 표적화되었다. 그러나, 고비율로 증가시키기 위한 포름알데히드의 효과와 관련하여 논쟁이 있었으므로 (문헌 [Chung, GTY, et al. 2005 Clin Chem 51, 655-658]), 방법의 재현성은 더 평가될 필요가 있다. 또한, 각각의 태아 및 모체는 각각의 염색체에 대해 상이한 수의 SNP의 정보를 제공하므로, SNP 비율 비교를 위한 통계적 시험의 영향력은 건 마다 변할 것이다 (문헌 [Lo, YMD & Chiu, RWK. 2007 Lancet 369, 1997]). 또한, 이들 접근법은 유전자 다형성의 검출에 좌우되므로, 이들 다형성에 이형접합성인 태아에게 국한된다.

[0013] 삼염색체 21 및 정배수체 태아로부터 수득한 양수세포 배양물에서 염색체 21 유전자좌 및 참조 유전자좌의 중합효소 연쇄 반응 (PCR) 및 DNA 정량화를 사용하여, 짐메르만 등(Zimmermann et al) (문헌 [2002 Clin Chem 48, 362-363])은 삼염색체 21 태아에서 염색체 21 DNA 서열의 1.5배 증가를 기준으로 상기 2군의 태아를 구별할 수 있었다. DNA 주형 농도의 2배 차이는 단지 한 임계 사이클(threshold cycle (Ct))의 차이를 구성하기 때문에, 1.5배 차이의 구별은 통상적인 실시간 PCR의 한계였다. 보다 미세한 정도의 정량 구별을 달성하기 위해, 별법의 전략이 필요하다. 따라서, 본 발명의 일부 실시양태는 이러한 목적을 위해 디지탈 PCR (문헌 [Vogelstein, B et al. 1999 Proc Natl Acad Sci USA 96, 9236-9241])을 사용한다.

[0014] 디지탈 PCR은 핵산 샘플에서 대립유전자 비율이 벗어난 것을 검출하기 위해 개발되었다 (문헌 [Chang, HW et al. 2002 J Natl Cancer Inst 94, 1697-1703]). 임상적으로, 디지탈 PCR은 종양 DNA 샘플에서 이형접합성의 소실 (LOH)을 검출하는데 유용한 것으로 나타났다 (문헌 [Zhou, W. et al. 2002 Lancet 359, 219-225]). 디지

탈 PCR 결과의 분석을 위해, 순차 확률비 검증 (SPRT)이 이전 연구에 의해 채택되어, 샘플 중 LOH의 존재를 제시하는 것과 제시하지 않는 것으로 실험 결과를 분류하였다 (문헌 [El Karoui at al. 2006 Stat Med 25, 3124-3133]). 이전 연구에서 사용된 방법에서, LOH를 결정하기 위한 분별값 (cutoff value)은 DNA 중 2개 대립유전자의 고정 참조 비율 2/3을 사용하였다. 모체 혈장 중 태아 핵산의 양, 비율 및 농도는 변할 수 있기 때문에, 이들 방법은 모체 혈장 중 모체 핵산의 배경에서 태아 핵산을 사용하여 삼염색체 21을 검출하는데 적합하지 않다.

[0015] 순환 태아 핵산 분석을 기준으로 하는 태아 삼염색체 21 (및 다른 불균형) 검출을 위한 비-침습적 시험, 특히 유전자 다형성 및/또는 태아-특이적 마커의 사용에 독립적인 비-침습적 시험을 수행하는 것이 바람직하다. 또한, 정확한 분별값 결정 및 서열 계수를 수행하여, 정확성에 필요한 모체 혈장 핵산 분자의 양 및/또는 데이타의 웰 수를 감소시킬 수 있음으로써, 증가된 효율 및 비용-효과를 제공하는 것이 바람직하다. 또한, 비-침습적 시험이 고 감수성 및 특이성을 가져 오류 진단을 최소화하는 것이 바람직하다.

[0016] 모체 혈장 중 태아 DNA 검출에 대한 또다른 적용은 베타-지중해빈혈과 같은 단일 유전자 장애의 출생 전 진단에 대한 것이다. 그러나, 태아 DNA는 단지 모체 혈장 중 소수 분획의 DNA를 구성하므로, 이러한 접근법은 단지 태아에서 부계로부터 유전되었지만 모계로부터는 존재하지 않는 돌연변이를 검출할 수 있을 뿐이라고 생각된다. 이러한 예로는, 베타-지중해빈혈을 일으키는 베타-글로빈 유전자의 코돈 41/42에서의 4 bp 결실 (문헌 [Chiu RWK et al. 2002 Lancet, 360, 998-1000]) 및 낭포성 섬유증을 일으키는 낭포성 섬유증 막횡단 전도 조절자 유전자의 Q890X 돌연변이 (문헌 [Gonzalez-Gonzalez et al 2002 Prenat Diagn, 22, 946-8])가 포함된다. 그러나, 베타-지중해빈혈 및 낭포성 섬유증 둘 다는, 이 질환이 그 자체로 표출되기 전에 태아에게 각각의 부모로부터 돌연변이가 유전될 필요가 있는 상염색체 열성 상태이므로, 부계-유전 돌연변이만의 검출은 단지 상기 질환에 걸린 태아를 가질 위험을 25%에서 50%로 증가시킬 것이다. 진단적으로 이는 이상적이 아니다. 따라서, 현존하는 접근법의 주된 진단 적용은, 부계-유전된 태아 돌연변이가 모체 혈장에서 검출될 수 없는 경우, 이후 태아가 동형접합성 질환 상태에 걸리는 것으로부터 배제될 수 있는 경우의 시나리오를 위해서일 것이다. 그러나, 진단적으로, 이러한 접근법은 부계 돌연변이의 음성 검출을 기준으로 결론을 내린다는 단점을 갖는다. 따라서, 상기 제한 없이, 완전한 태아 유전자형 (동형접합성 정상, 동형접합성 돌연변이 또는 이형접합성)이 모체 혈장으로부터 결정되도록 하는 접근법이 매우 바람직할 것이다.

발명의 내용

해결하려는 과제

[0018] 개요

[0019] 본 발명의 실시양태는, 핵산 서열 불균형 (예를 들어, 대립유전자 불균형, 돌연변이 불균형 또는 염색체 불균형)이 생물학적 샘플 내에 존재하는지 여부를 결정하기 위한 방법, 시스템 및 장치를 제공한다. 예를 들어, 2개의 서열 (또는 서열 세트) 양의 비율의 불균형을 결정하기 위한 하나 이상의 분별값이 선택된다.

[0020] 한 실시양태에서, 분별값은 적어도 부분적으로 생물학적 샘플, 예컨대 모체 핵산 서열의 배경을 함유하는 모체 혈장 또는 혈청 또는 뇨 중 태아 (임상적으로 관련된 핵산) 서열의 백분율을 기준으로 결정된다. 또다른 실시양태에서, 분별값은 다수의 반응에서 서열의 평균 농도를 기준으로 결정된다. 한 측면에서, 분별값은 특징 핵산 서열을 함유하는 것으로 평가되는 정보제공 웰의 비율로부터 결정되며, 그 비율은 상기 언급된 백분율 및/또는 평균 농도를 기준으로 결정된다.

[0021] 분별값은 다수의 상이한 유형의 방법, 예컨대 SPRT, 오류 발견, 신뢰 구간, 수신자 판단 특성 (receiver operating characteristic, ROC)을 사용하여 결정될 수 있다. 이러한 전략은 신뢰성 있는 분류가 만들어질 수 있기 전에 필요한 시험의 양을 더 최소화시켰다. 이는 주형 양이 종종 제한적인 혈장 핵산 분석과 특히 관련되어 있다.

[0022] 한 예시적 실시양태에 따라, (1) 임상적으로 관련된 핵산 서열의 제1 양을 나타내는 제1 세트의 정량 데이타; 및 (2) 임상적으로 관련된 핵산 서열과 상이한 배경 핵산 서열의 제2 양을 나타내는 제2 세트의 정량 데이타를 포함하는 데이타를 다수의 반응으로부터 수득하는 단계; 두 데이타 세트로부터 파라미터를 결정하는 단계; 다수의 반응 각각에서, 임상적으로 관련된 핵산 서열 또는 배경 핵산 서열인 참조 핵산 서열의 평균 농도로부터 제1 분별값을 유도하는 단계; 파라미터를 제1 분별값과 비교하는 단계; 및 비교를 기준으로, 핵산 서열 불균형이 존재하는지 여부의 분류를 결정하는 단계를 포함하는, 핵산 서열 불균형이 생물학적 샘플 내에 존재하는지 여부를

결정하기 위한 방법이 제공된다.

[0023] 또다른 예시적 실시양태에 따라, (1) 임상적으로 관련된 핵산 서열의 제1 양을 나타내는 제1 세트의 정량 데이타; 및 (2) 임상적으로 관련된 핵산 서열과 상이한 배경 핵산 서열의 제2 양을 나타내는 제2 세트의 정량 데이타를 포함하며, 임상적으로 관련된 핵산 서열 및 배경 핵산 서열은 제1 유형 세포 및 하나 이상의 제2 유형 세포로부터 유래된 것인 데이타를 다수의 반응으로부터 수득하는 단계; 두 데이타 세트로부터 파라미터를 결정하는 단계; 생물학적 샘플에서, 제1 유형 세포로부터의 핵산 서열의 양을 측정하여 생성된 제1 백분율로부터 제1 분별값을 유도하는 단계; 파라미터를 분별값과 비교하는 단계; 및 비교를 기준으로, 핵산 서열 불균형이 존재하는지 여부의 분류를 결정하는 단계를 포함하는, 핵산 서열 불균형이 생물학적 샘플 내에 존재하는지 여부를 결정하기 위한 방법이 제공된다.

[0024] 본 발명의 다른 실시양태는 본원에 기재된 방법과 관련된 시스템 및 컴퓨터 판독가능 매체에 관한 것이다. 예시적 실시양태에 따르면, 본 발명은 핵산 서열 불균형이 생물학적 샘플 내에 존재하는지 여부를 결정하기 위한 작업을 수행하기 위한 컴퓨터 시스템 제어용의 다수의 지시로 코딩된, 비일시적 컴퓨터 판독가능 기록 매체에 관한 것으로서, 상기 지시는 (1) 생물학적 샘플 중의 임상적으로 관련된 핵산 서열의 제1 양을 나타내는 제1 세트의 정량 데이터; 및 (2) 임상적으로 관련된 핵산 서열과 상이한 생물학적 샘플 중의 배경 핵산 서열의 제2 양을 나타내는 제2 세트의 정량 데이터를 포함하고, 상기 임상적으로 관련된 핵산 서열과 배경 핵산 서열이 제1 유형의 세포 및 하나 이상의 제2 유형의 세포로부터 유래된 것인, 다수의 반응으로부터 데이터를 수득하는 단계; 두 데이터 세트로부터 파라미터를 결정하는 단계; 생물학적 샘플에서 제1 유형의 세포로부터의 핵산 서열의 양을 측정하여 수득된 제1 백분율로부터 제1 분별값을 유도하는 단계; 파라미터를 제1 분별값과 비교하는 단계; 및 비교를 기준으로, 핵산 서열 불균형이 존재하는지 여부의 분류를 결정하는 단계를 포함한다.

[0025] 하기 상세한 설명 및 수반되는 도면을 참조로 하여 본 발명의 성질 및 이점을 보다 잘 이해할 수 있다.

발명의 실시 방법

[0027] 정의

[0028] 본원에 사용된 용어 "생물학적 샘플"은 대상체 (예를 들어, 임신 여성과 같은 인간)로부터 취하며 하나 이상의 관심 핵산 분자(들)을 함유하는 임의의 샘플을 지칭한다.

[0029] 용어 "핵산" 또는 "폴리뉴클레오티드" 단일가닥 또는 이중가닥 형태의 데옥시리보핵산 (DNA) 또는 리보핵산 (RNA) 및 이들의 중합체를 지칭한다. 구체적으로 한정하지 않는 한, 상기 용어는 참조 핵산과 유사한 결합 성질을 가지며 천연 발생 뉴클레오티드와 유사한 방식으로 대사되는 천연 뉴클레오티드의 공지된 유사체를 함유하는 핵산을 포함한다. 달리 나타내지 않는 한, 특정 핵산 서열은 또한 암묵적으로, 그의 보존적으로 변형된 변이체 (예를 들어, 퇴행 코돈 치환체), 대립유전자, 오소로그(ortholog), SNP 및 상보적 서열, 및 명백하게 나타내어진 서열을 포함한다. 구체적으로, 퇴행 코돈 치환체는 하나 이상의 선택된 (또는 모든) 코돈의 제3 위치가 혼합 염기 및/또는 데옥시이노신 잔기로 치환되는 서열을 생성함으로써 달성될 수 있다 (문헌 [Batzer et al, Nucleic Acid Res. 19:5081 (1991)], [Ohtsuka et al., J. Biol. Chem. 260:2605-2608 (1985)] 및 [Rossolini et al., Mol. Cell. Probes 8:91-98 (1994)]). 용어 핵산은 유전자, cDNA, mRNA, 소형 비코딩 RNA, 마이크로 RNA (miRNA), 피위(Piwi)-상호작용 RNA, 및 유전자 또는 유전자좌에 의해 코딩되는 단형 헤어핀 RNA (shRNA)와 상호교환적으로 사용된다.

[0030] 용어 "유전자"는 폴리펩티드 쇄를 생성하는데 관여하는 DNA 절편을 의미한다. 코딩 영역 전 및 후 영역 (리더 및 트레일러) 뿐만 아니라 개별 코딩 절편 (엑손) 사이의 삽입 서열 (인트론)을 포함할 수 있다.

[0031] 본원에 사용된 용어 "반응"은 관심있는 특정 폴리뉴클레오티드 서열의 존재 또는 부재를 나타내는 화학적, 효소적 또는 물리적 작용에 관여하는 임의의 과정을 지칭한다. "반응"의 한 예는 중합효소 연쇄 반응 (PCR)과 같은 증폭 반응이다. "반응"의 또다른 예는 합성 또는 라이게이션에 의한 서열분석 반응이다. "정보제공 반응"은 관심있는 하나 이상의 특정 폴리뉴클레오티드 서열의 존재를 나타내며 단지 하나의 관심 서열이 존재하는 한가지 경우에서의 반응이다. 본원에 사용된 용어 "웰"은 제한된 구조, 예를 들어 웰-형상 바이알, 셀, 또는 PCR 어레이 챔버 내의 예정된 위치에서의 반응을 지칭한다.

[0032] 본원에 사용된 용어 "임상적으로 관련된 핵산 서열"은 잠재적 불균형에 대해 시험될 보다 큰 게놈 서열의 절편에 상응하는 폴리뉴클레오티드 서열 또는 보다 큰 게놈 서열 그 자체를 지칭할 수 있다. 한 예는 염색체 21의 서열이다. 다른 예는 염색체 18, 13, X 및 Y를 포함한다. 또다른 예는 돌연변이 유전자 서열 또는 유전자 다

형성 또는 부모의 한쪽 또는 양쪽으로부터 태아에게 유전될 수 있는 카피 수 변형을 포함한다. 또다른 예는 악성 종양에서 돌연변이, 결실 또는 증폭된 서열, 예를 들어 이형접합성 소실 또는 유전자 복제가 일어나는 서열을 포함한다. 일부 실시양태에서, 다수의 임상적으로 관련된 핵산 서열 또는 동등하게는 임상적으로 관련된 핵산 서열의 다수의 마커는 불균형 검출을 위한 데이타를 제공하는데 사용될 수 있다. 예를 들어, 염색체 21 상의 5개의 비-연속 서열로부터의 데이타는 가능한 염색체 21 불균형의 결정을 위해 추가적인 방식으로 사용되어, 필요한 샘플 부피를 1/5로 유효하게 감소시킬 수 있다.

[0033] 본원에 사용된 용어 "배경 핵산 서열"은 임상적으로 관련된 핵산 서열에 대한 정상 비율이 공지되어 있는 (예를 들어 1 대 1 비율인) 핵산 서열을 지칭한다. 한 예로서, 배경 핵산 서열 및 임상적으로 관련된 핵산 서열은 이형접합성으로 인해 구별되는, 동일한 염색체로부터의 2개의 대립유전자이다. 또다른 예로, 배경 핵산 서열은 임상적으로 관련된 핵산 서열인 다른 대립유전자에 대해 이형접합성인 한 대립유전자이다. 또한, 각각의 배경 핵산 서열 및 임상적으로 관련된 핵산 서열 중 일부는 상이한 개체로부터 유래될 수 있다.

[0034] 본원에 사용된 용어 "참조 핵산 서열"은 반응 당 평균 농도가 공지되어 있거나 동등하게는 측정되는 핵산 서열을 지칭한다.

[0035] 본원에 사용된 용어 "과나표출된(overrepresented) 핵산 서열"은 2개의 관심 서열 (예를 들어, 임상적으로 관련된 서열 및 배경 서열) 중 생물학적 샘플에서 다른 서열보다 풍부한 핵산 서열을 지칭한다.

[0036] 본원에 사용된 용어 "~기준으로"는 "적어도 부분적으로 ~기준으로"를 의미하며, 방법의 투입과 방법의 산출의 관계에서 일어나는, 한 값 (또는 결과)이 또다른 값의 결정에 사용되는 것을 지칭한다. 본원에 사용된 용어 "유도"는 그 유도가 식의 계산인 경우에 일어나는, 방법의 투입과 방법의 산출의 관계를 지칭한다.

[0037] 본원에 사용된 용어 "정량 데이타"는 하나 이상의 수치 값을 제공하는 하나 이상의 반응으로부터 얻은 데이타를 의미한다. 예를 들어, 특정 서열에 대한 형광 마커를 나타내는 웰의 수는 정량 데이타일 것이다.

[0038] 본원에 사용된 용어 "파라미터"는 정량 데이타 세트, 및/또는 정량 데이타 세트들 사이의 수치 관계를 특성화하는 수치 값을 의미한다. 예를 들어, 제1 핵산 서열의 제1 양과 제2 핵산 서열의 제2 양 사이의 비율 (또는 비율의 함수)이 파라미터이다.

[0039] 본원에 사용된 용어 "분별값"은 생물학적 샘플에 대한 분류의 2가지 이상의 상태 (예를 들어 질환 상태 및 비질환 상태)를 증재하는데 사용되는 수치 값을 의미한다. 예를 들어, 파라미터가 분별값보다 큰 경우 정량 데이타의 제1 분류가 만들어지거나 (예를 들어 질환 상태), 또는 파라미터가 분별값보다 작은 경우 정량 데이타의 상이한 분류가 만들어진다 (예를 들어 비질환 상태).

[0040] 본원에 사용된 용어 "불균형"은 하나 이상의 분별값으로 정의되는, 참조 정량으로부터 임상적으로 관련된 핵산 서열의 정량의 임의의 유의한 편차를 의미한다. 예를 들어, 참조 정량이 3/5의 비율일 수 있고, 따라서 측정 비율이 1:1인 경우에 불균형이 일어날 것이다.

[0042] 발명의 상세한 설명

[0043] 본 발명은, 다른 비-임상적으로 관련된 서열 (예를 들어, 염색체 또는 대립유전자 불균형)에 관하여 임상적으로 관련된 핵산 서열의 참조 (예를 들어 비질환 상태) 양에 비교되는 증가 또는 감소가 생물학적 샘플 내에 존재하는지 여부를 결정하기 위한 방법, 시스템 및 장치를 제공한다. 예를 들어 2개의 서열 (또는 서열 세트) 양의 비율과 관련하여 참조 양에 비교되는 변화 (즉 불균형)이 존재하는지 여부를 결정하기 위해 하나 이상의 분별값이 선택된다. 참조 양에서 검출된 변화는 다른 비-임상적 관련 서열에 대한 임상적 관련 핵산 서열에 관하여 임의의 편차 (상향 또는 하향)일 수 있다. 따라서, 참조 상태는 임의의 비율 또는 다른 양 (예를 들어 1-1 대응 이외의 것)일 수 있고, 변화를 의미하는 측정 상태는 하나 이상의 분별값에 의해 결정된 바와 같이 참조 양과 상이한 다른 양 또는 임의의 비율일 수 있다.

[0044] 임상적으로 관련된 핵산 서열 및 배경 핵산 서열은 제1 유형 세포 및 하나 이상의 제2 유형 세포로부터 유래될 수 있다. 예를 들어, 태아/태반 세포로부터 유래된 태아 핵산 서열은 생물학적 샘플, 예컨대 모체 세포로부터 유래된 모체 핵산 서열의 배경을 함유하는 모체 혈장 중에 존재한다. 따라서, 한 실시양태에서, 분별값은 적어도 부분적으로, 생물학적 샘플 중 제1 유형 세포의 백분율을 기준으로 결정된다. 샘플 중 태아 서열의 백분율은 임의의 태아-유래 유전자좌에 의해 결정될 수 있으며, 임상적으로 관련된 핵산 서열을 측정하는 것에 국한되지 않는다는 것을 주목한다. 또다른 실시양태에서, 분별값은 적어도 부분적으로, 생물학적 샘플, 예컨대 체내 비-악성 세포로부터 유래된 핵산 서열의 배경을 함유하는 혈장, 혈청, 타액 또는 뇨 중 종양 서열의 백분율을

기준으로 결정된다.

[0045] 또다른 실시양태에서, 분별값은 다수의 반응에서 서열의 평균 농도를 기준으로 결정된다. 한 측면에서, 분별값은 특정 핵산 서열을 함유하는 것으로 평가되는 정보제공 웰의 비율로부터 결정되며, 그 비율은 상기 언급된 백분율 및/또는 평균 농도를 기준으로 결정된다. 분별값은 다수의 상이한 유형의 방법, 예컨대 SPRT, 오류 발견, 신뢰 구간, 수신자 판단 특성 (ROC)을 사용하여 결정될 수 있다. 이러한 전략은 신뢰성 있는 분류가 만들어질 수 있기 전에 필요한 시험의 양을 더 최소화시킨다. 이는 주형 양이 종종 제한적인 혈장 핵산 분석과 특히 관련되어 있다. 디지탈 PCR이 제시되더라도, 다른 방법이 사용될 수 있다.

[0046] 디지탈 PCR은 대부분의 양성 증폭이 단일 주형 분자로부터의 신호를 반영하도록 매우 희석된 핵산에 대한 멀티플렉스 PCR 분석을 포함한다. 이로써 디지탈 PCR은 개별 주형 분자의 계수를 가능하게 한다. 분석된 PCR의 총 수 중 양성 증폭의 비율로 원래 또는 비-희석 샘플 중의 주형 농도를 평가할 수 있다. 이러한 기술은 다양한 유전 현상의 검출을 가능하게 하도록 제시되었으며 (상기 문헌 [Vogelstein, B et al. 1999]), 종양 샘플 중 이형접합성 소실의 검출 (상기 문헌 [Zhou, W. et al. 2002]) 및 암 환자의 혈장 중 이형접합성 소실의 검출 (상기 문헌 [Chang, HW et al. 2002])을 위해 이전부터 사용되었다. 디지탈 PCR에 의한 주형 분자 정량화는 리포터 염료와 핵산 농도 사이의 투여량-반응 관계에 의존하지 않기 때문에, 그의 분석 정확도는 실시간 PCR의 분석 정확도보다 이론적으로 월등하여야 한다. 따라서, 디지탈 PCR은 표적 유전자좌와 참조 유전자좌 사이의 보다 미세한 정도의 정량 차이를 잠재적으로 구별할 수 있게 한다.

[0047] 이를 시험하기 위해, 본 발명자들은 먼저 디지탈 PCR이 모체 혈장 중 염색체 21로부터의 태반 전사체인 PLAC4 mRNA (문헌 [Lo, YMD, et al. 2007 Nat Med 13, 218-223]) 대립유전자 비율을 결정하고, 이로써 삼염색체 21 및 정배수체 태아를 구별할 수 있는지 여부를 평가하였다. 이러한 접근법을 디지탈 RNA-SNP 방법이라 칭한다. 이어서, 본 발명자들은 디지탈 PCR의 증가된 정확도로 인해 유전자 다형성에 좌우되지 않으면서 태아 염색체 이수성을 검출할 수 있는지 여부를 평가하였다. 본 발명자들은 이를 디지탈 상대적 염색체 투여 (RCD) 분석이라 칭한다. 전자의 접근법은 다형성-의존적이지만 정량 구별에 있어서 보다 낮은 정확도를 필요로 하는 반면, 후자의 접근법은 다형성-독립적이지만 정량 구별에 있어서 보다 높은 정확도를 필요로 한다.

[0049] I. 디지탈 RNA-SNP

[0050] **A. 개관**

[0051] 디지탈 PCR은 DNA 샘플 중 두 대립유전자의 벗어난(skewing) 대립유전자 비율의 존재를 검출할 수 있다. 예를 들어, 디지탈 PCR은 종양 DNA 샘플 중 이형접합성 소실 (LOH)을 검출하는데 사용되어 왔다. DNA 샘플 중 두 대립유전자, 즉 A 및 G가 존재하고, A 대립유전자는 LOH가 있는 세포 중에 소실된다고 가정한다. LOH가 종양 샘플 중 50%의 세포에 존재하는 경우, DNA 샘플 중 G:A의 대립유전자 비율은 2:1일 것이다. 그러나, LOH가 종양 샘플에 존재하지 않는 경우, G:A의 대립유전자 비율은 1:1일 것이다.

[0052] 도 1은 디지탈 PCR 실험을 설명하는 흐름도 100이다. 단계 110에서, DNA 샘플을 희석시킨 후, 독립된 웰로 분배한다. 본 발명자들은 일부 혈장 핵산 종이 원래의 샘플에서 이미 상당히 희석되도록 결정했다는 것을 주목한다. 따라서, 필요한 농도로 이미 존재하는 경우, 일부 주형에 대한 희석은 필요하지 않다. 이전의 연구 (예를 들어, 상기 문헌 [Zhou et al 2002])에서는, 특정 "주형 DNA"의 평균 농도가 웰 당 두 주형 중 하나의 대략 0.5 분자가 될 정도로 DNA 샘플을 희석시켰다. 용어 "주형 DNA"는 A 또는 G 대립유전자를 가리키는 것으로 보이고, 이러한 특정 농도에 대한 이론적 근거는 제공되지 않았다는 것을 주목한다.

[0053] 단계 120에서, 각각의 웰에서, PCR 과정은 A 및/또는 G 대립유전자를 동시에 검출하기 위해서 실시된다. 단계 130에서, 각각의 웰 중 마커가 식별된다 (예를 들어, 형광으로) (예를 들어 A, G, A 및 G, 또는 어느 쪽도 아님). LOH의 부재 하에, DNA 샘플 중 A 및 G 대립유전자의 과다는 동일하다 (셀 당 각각 한 카피). 따라서, 웰이 A 대립유전자 및 G 대립유전자에 대해 양성일 확률은 동일할 것이다. 이것은 유사한 수의 웰이 A 또는 G 대립유전자에 대해 양성인 것에 의해 반영될 것이다. 그러나, LOH가 종양 샘플 중 50% 또는 그 보다 많은 세포에 존재하는 경우, G 및 A 대립유전자의 대립유전자 비율은 적어도 2:1일 것이다. 이전의 방법들은 단순히 샘플이 적어도 50% 암성인 것으로 추정하였다. 따라서, 웰이 G 대립유전자에 대해 양성일 확률은 A 대립유전자에 대해 양성일 확률보다 높을 것이다. 그 결과, G 대립유전자에 대해 양성인 웰의 수는 A 대립유전자에 대해 양성인 웰의 수보다 더 많을 것이다.

[0054] 단계 140에서, 디지탈 PCR 결과를 분류하기 위해, 각각의 대립유전자에 대해 양성이나 다른 대립유전자에 대해서는 양성이 아닌 웰의 수를 계수할 것이다. 상기 예에서, A 대립유전자에 대해 양성이나 G 대립유전자에 대해

음성인 웰의 수, 및 G 대립유전자에 대해 양성이나 A 대립유전자에 대해 음성인 웰의 수를 계수한다. 한 실시양태에서, 딘 양성인 웰을 나타내는 대립유전자는 참조 대립유전자로 간주된다.

[0055] 단계 150에서, 정보제공 웰의 총 수는 두 대립유전자 중 어느 하나에 대해 양성인 웰의 수의 합으로 결정된다. 단계 160에서, 더 많은 양성 웰의 대립유전자에 의해 기여되는 정보제공 웰의 비율 (P_r) (파라미터의 한 예)이 계산된다.

[0056] P_r = 더 많은 양성 웰의 대립유전자에 대해서만 양성인 웰의 수/단지 하나의 대립유전자 (A 또는 G)에 대해 양성인 웰의 총 수.

[0057] 다른 실시양태는 하나의 대립유전자를 갖는 모든 웰을 하나 이상의 대립유전자를 갖는 모든 웰로 나눈 것을 사용할 수 있다.

[0058] 단계 170에서, P_r 값이 대립유전자 불균형을 나타내는지 여부가 결정된다. 정확하고 효율적인 것이 바람직하므로, 이 작업은 간단하지 않다. 불균형을 결정하기 위한 한 방법은 베이지안(Bayesian)-유형 우도 방법, 순차 확률비 검증 (SPRT)을 사용한다. SPRT는 데이타가 축적되면서 두 확률 가설을 비교하도록 하는 방법이다. 즉, SPRT는 디지탈 PCR의 결과를 분류하여 벗어난 대립유전자의 존재 또는 부재를 시사하는 통계적 방법이다. SPRT는 분석될 웰의 수를 최소화하여 소정의 통계적 검증력 및 정확도를 획득하도록 하는 이점을 갖는다.

[0059] 예시적 SPRT 분석에서, 실험 결과는 귀무가설 (null hypothesis) 및 대립가설 (alternative hypothesis)에 대해 시험될 것이다. 대립가설은 샘플 중 벗어난 대립유전자 비율이 있는 경우 수용된다. 귀무가설은 샘플 중 벗어난 대립유전자 비율이 없는 경우 수용된다. P_r 값은 귀무가설 또는 대립가설을 수용하기 위해서 두 분별값과 비교될 것이다. 두 가설 모두 수용되지 않는 경우, 샘플은 관찰된 디지탈 PCR 결과가 바람직한 통계적 신뢰도로 샘플을 분류하기에 충분하지 않다는 것을 의미하는 미분류로 표시될 것이다.

[0060] 귀무가설 또는 대립가설을 수용하기 위한 분별값은 전형적으로, 가설에서 언급된 가정 하에 고정된 P_r 값을 기준으로 계산되어 왔다. 귀무가설에서, 샘플은 벗어난 대립유전자 비율을 나타내지 않는 것으로 가정된다. 따라서, 각 웰이 A 및 G 대립유전자에 대해 양성일 확률은 동일할 것이므로, 예상 P_r 값은 1/2일 것이다. 대립가설에서, 예상 P_r 값은 2/3 또는 0.5와 2/3의 대략 중간 정도, 예를 들어 0.585로 여겨진다. 또한, 제한된 수의 실험으로 인해, (.585+3/N)의 상한값 및 (.585-3/N)의 하한값이 선택될 수 있다.

[0062] **B. 다운 증후군의 검출**

[0063] 본 발명의 한 실시양태에서, 디지탈 SNP는 임신 여성의 혈장으로부터 태아 다운 증후군을 검출하는데 사용된다. 태아/태반 세포에 특이적인 마커를 사용하여, 염색체 21 중 대립유전자의 비율을 측정할 수 있다. 예를 들어, PLAC4 대립유전자의 관찰된 과다표출 정도가 통계적으로 유의한지를 결정하기 위해서, SPRT가 사용된다.

[0064] 하나의 예시적 실시양태에 따라, 디지탈 RNA-SNP는 염색체 21로부터 전사되고 태반에 의해 발현되는 PLAC4 mRNA에 위치한 A/G SNP의 다형성 대립유전자, rs8130833의 비율에서의 불균형을 결정한다. 이형접합 정배수체 태아에 대해, A 및 G 대립유전자는 태아 게놈에서 동등하게 표출되어야 하는 한편 (1:1 게놈 비율); 삼염색체 21에서, 삼염색체성 염색체 21은 태아 게놈 중 하나의 SNP 대립유전자의 추가 카피와 회합되어 2:1 비율을 나타낼 것이다. 디지탈 PCR 분석의 목적은 분석된 샘플 중 두 PLAC4 대립유전자의 양이 동등한지 또는 다른지를 결정하는 것이다. 따라서, A 및 G PLAC4 대립유전자는 모두 표적 주형이다. 실시간 PCR 검정은 PLAC4 mRNA를 증폭시키도록 설계되었고, 두 SNP 대립유전자는 태크만(TaqMan) 형광 프로브에 의해 구별되었다. 분석 단계의 도식적 설명이 도 2a에 나타나 있다.

[0065] 도 2a는 본 발명의 실시양태에 따른 디지탈 RNA-SNP 방법 200을 설명한다. 단계 210에서, 샘플을 얻는다. 단계 220에서, 핵산 서열, 예를 들어 PLAC4 mRNA를 추출된 RNA 샘플에서 정량화한다. 한 실시양태에서, 이것은 PLAC4 mRNA에 대해 실시간 PCR로 실시된다. 한 측면에서, 이 단계는 조작자에게 표적이 디지탈 PCR 분석의 '영역'에 도달하기 전에 어느 만큼의 희석이 요구되는지에 대한 아이디어를 제공한다.

[0066] 단계 230에서, 샘플을 희석시킨다. 단계 240에서, 희석시킨 샘플의 농도를 측정한다. 희석시킨 샘플의 농도가 약 1 주형/웰 (즉, 참조 또는 비-참조 서열 또는 어느 하나의 대립유전자)인지 확인할 수 있다. 일부 실시양태는 이 측정을 위한 IV절에 기술된 기술을 사용한다. 예를 들어, 본 발명자들은 실시간 PCR 분석을 위해서 희석시킨 샘플을 96 웰로 배분하여 알맞게 희석되었는지를 확인하였다. 희석 농도는 또한 모르는 상태로 남겨두어

이 단계가 제거되도록 할 수 있고, 이는 추후 설명될 것이다.

[0067] 단계 250에서, 디지탈 PCR을 어레이의 각 웰 상에서 실시하였다. 예를 들어, 동일하게 희석시킨 샘플을 실시간 PCR 분석을 위해서 384-웰에 배분하였다. PCR 결과로부터, 각 핵산 서열에 대한 마커의 양 및 정보제공 웰의 수를 확인한다. 정보제공 웰은 A 또는 G 대립유전자에 대해서만 양성이나, 둘 모두에 대해서 양성은 아닌 것으로 정의된다. 단계 260에서, 예상 P_r 값을 계산한다. 이러한 단계들은 추후 더욱 자세하게 설명될 것이다. 이 계산은 단계 250에서 측정된 값들로부터 파라미터를 결정하는 것을 포함한다. 예를 들어, 웰 당 실제 평균 주형 농도가 계산될 수 있다.

[0068] 단계 270에서, SPRT 또는 다른 우도-비율 시험은 불균형이 존재하는지 여부를 결정하기 위해서 실시될 수 있다. 정배수체 사례에 대해, 동일한 수의 A-양성 및 G-양성 웰이 예상된다. 그러나, 삼염색체 21 태아로부터의 주형 분자가 분석되는 경우, 하나의 대립유전자만을 함유하는 웰의 수는 다른 대립유전자만을 함유하는 웰의 수보다 많아야 한다. 요컨대, 삼염색체 21에 대한 대립유전자 불균형이 예상된다.

[0069] 상기 언급된 것과 같이, SPRT는 데이타가 축적되면서 두 확률 가설을 비교하도록 하는 베이지안-유형 우도 방법이다. 삼염색체 21 검출을 위한 디지탈 PCR 분석에서, 대립가설은 대립유전자 불균형이 존재하는 경우 (즉, 삼염색체 21이 검출되는 경우) 수용되고; 귀무가설은 대립유전자 불균형이 존재하지 않는 경우 (즉, 삼염색체 21이 검출되지 않는 경우) 수용된다. 더 많은 수의 계수의 대립유전자는 잠재적으로 과다표출된 대립유전자로 지정되고, 모든 정보제공 웰 중 그의 비율 (P_r)이 계산될 것이다. SPRT는, P_r이 삼염색체 21 샘플에 대해 예상되는 충분한 정도의 대립유전자 불균형을 나타내는지를 결정하기 위해서 적용된다.

[0070] 조작 면에 있어서, SPRT는 어느 하나의 가설을 수용하거나 또는 거절하기 위한 확률 경계를 정의하기 위해서 구축된 한 쌍의 SPRT 곡선을 갖는 그래프를 사용하여 적용되고 해석될 수 있다. 도 3은 본 발명의 한 실시양태에 따라 다운 증후군을 결정하기 위한 SPRT 곡선을 갖는 그레프를 설명한다. SPRT 곡선은 신뢰성 있게 분류될 수 있는 경우 주어진 정보제공 웰의 총 수 (x-축)에 대해 잠재적으로 과다표출된 대립유전자에 대해 양성인 정보제공 웰의 요구되는 비율 P_r (y-축)을 플롯팅한다. 도 3에 설명된 것과 같이, 상단 곡선은 대립가설을 수용하기 위한 확률 경계를 설정하고; 하단 곡선은 귀무가설을 수용하기 위한 확률 경계를 설정한다.

[0071] 실험적으로 유도된 P_r 값은 어느 하나의 가설을 수용하거나 또는 거절하기 위한 예상 P_r 값과 비교될 것이다. 귀무가설이 수용되는 경우, 샘플은 정배수체 태아를 가진 임신 여성으로부터 수득된 것으로 분류되었다. 대립가설이 수용되는 경우, 샘플은 삼염색체 21 태아를 가진 임신 여성으로부터 수득된 것으로 분류되었다. 별법으로, 주어진 정보제공 계수의 수에 대한 P_r이 질환 분류에 대해 요구되는 통계적 신뢰성 수준에 아직 도달하지 못한 경우, 어느 하나의 가설도 받아들여질 수 없다. 이러한 경우들은 더 많은 데이타가 이용가능하기 전까지는 분류불가능으로 간주된다. 질환 분류가 가능하지 않은 경우, 집계된 데이타가 SPRT에 의해 분류가능해지기 전까지 추가의 384-웰 플레이트를 실시할 수 있다.

[0072] 따라서, SPRT는 주어진 수준의 신뢰성을 위해서 다른 통계적 방법보다 더 적은 양의 시험이 요구된다는 이점을 제공한다. 실용적인 측면에서, SPRT는 요구되는 양의 데이타가 축적되어 불필요한 추가적 분석을 최소화하자마자 어느 하나의 가설의 수용 또는 거절을 가능하게 한다. 이 특징은 일반적으로 이용가능한 주형 분자의 수가 제한적인 경우 낮은 농도로 존재하는 혈장 핵산의 분석과 특히 관련이 있다. 엄격한 분류에 추가하여, 분류는 또한 백분율 정확성을 포함할 수 있다. 예를 들어, 분별값과의 비교로 인한 분류는 샘플이 특정 백분율의 핵산 서열 불균형의 우도를 나타내거나, 또는 동등하게는 측정된 불균형이 특정 백분율 또는 다른 값에 정확하다는 것을 제공할 수 있다.

[0073] 유사한 접근법을 적용하여, 모체 혈장 또는 혈청 중 태아 핵산을 사용하여 돌연변이 또는 유전자 다형성과 관련한 태아의 유전자형을 결정할 수 있다. 태아는 그의 모계로부터 그의 게놈의 절반을 물려받을 것이라는 것을 상기해야 한다. 설명을 위해서, 두 대립유전자 A 및 B가 있는 특정 유전자좌를 고려한다. 모계가 AB 유전자형의 이형접합체인 경우, 태아는 이론적으로는 AA, BB 또는 AB의 유전자형을 가질 수 있다. 태아가 AB 유전자형, 즉 모계와 같은 유전자형을 갖는 경우, 모체 혈장에는 (모계 및 태아 모두로부터의) AB 유전자형의 핵산만 존재할 것이다. 따라서, 핵산 또는 대립유전자 균형이 모체 혈장에서 보여진다. 반면, 태아가 AA 또는 BB의 유전자형을 갖는 경우, 모체 혈장에서 A 또는 B 대립유전자 각각이 과다표출된 대립유전자 불균형이 존재할 것이다. 이러한 생각은 또한 질환-유발성 돌연변이 (예를 들어, 낭포성 섬유증 또는 베타-지중해빈혈 또는 척수 근위축증을 유발하는 돌연변이)에도 적용가능한데, 각 경우 A는 야생형 대립유전자로, B는 돌연변이 대립유전자로 고

려될 수 있다.

[0075] II. 디지탈 RCD

[0076] 디지탈 RNA-SNP의 단점은 디지탈 RNA-SNP가 분석된 SNP에 대한 이형접합체의 경우에만 적용가능하다는 것이다. 순환성 태아 핵산 분석을 기준으로 한 삼염색체 21 태아 또는 다른 태아 염색체 이수성 (예를 들어, 삼염색체 18, 13 및 성 염색체 이수성)을 검출하기 위한 비침습성 시험이 유전자 다형성의 사용과 무관한 경우 디지탈 RNA-SNP가 이상적이라는 것은 하나의 개선된 점이다. 따라서, 한 실시양태에서, 염색체 양은 참조 염색체, 즉 이 연구에서 염색체 1에 위치한 유전자좌에 상대적인 비-다형성 염색체 21 유전자좌의 디지탈 PCR 분석에 의해 결정된다. 정배수체 태아의 게놈에서 염색체 21 대 염색체 1의 비율의 2:2로부터의 변화는 삼염색체 21의 경우와 구분된다. 삼염색체 21 검출을 위한 디지탈 PCR 분석에서, 비교할 두 가설은 염색체 불균형이 존재하지 않는 (즉, 삼염색체 21이 검출되지 않는) 귀무가설 및 염색체 불균형이 존재하는 (즉, 삼염색체 21이 검출되는) 대립가설일 것이다.

[0077] 이 접근법은 다른 염색체 이수성, 예를 들어 삼염색체 18에서의 염색체 18, 삼염색체 13에서의 염색체 13, 터너 증후군에서의 염색체 X와 연관된 다른 염색체에 일반화될 수 있다. 또한, 염색체 1과는 별개로, 관련 이수성과 연관되지 않은 다른 염색체도 또한 참조 염색체로 사용될 수 있다. 유사한 접근법이 또한 부분적으로, 암에서 통상적으로 결실되는 염색체 대 참조 염색체의 비율 변화를 분석함으로써 암 검출에 적용될 수 있다. 전자의 예로는 길장직장암에서의 염색체 5q, 폐암에서의 염색체 3p 및 비인두 암종에서의 염색체 9p가 포함된다. 도 2b는 서열 불균형으로 나타나는 일부 통상적인 암-관련 염색체 이상을 열거한다.

[0078] 도 2a는 또한 본 발명의 한 실시양태에 따른 디지탈 RCD 방법 205를 설명한다. 단계 220-230에 대한 한 실시양태에서, 추출된 DNA를, 예를 들어, 나노드롭(Nanodrop) 기술로 정량화하고, 웰 당 염색체 21 또는 표준화 염색체 (예컨대, 염색체 1)로부터의 약 하나의 표적 주형의 농도로 희석시킨다. 단계 240의 한 실시양태에서, 384-웰 플레이트 중 태크만 프로브를 모두 사용하는 디지탈 RCD 분석을 진행하기 전에 약 37% 수준의 웰이 음성인지 여부를 확인하기 위해서, 단지 96-웰 포맷에서 염색체 1 프로브를 사용하는 검정에 의해 희석시킨 DNA 샘플을 분석함으로써 확인할 수 있다. 37%의 유의성은 IV절에서 추후 기술될 것이다.

[0079] 단계 240의 시험 및 단계 250의 결과는 한 쌍의 태크만 프로프에 의해 구별되는 파라로그(paralogous) 서열 변이에 의해 식별되는 두 염색체 모두에 존재하는 파라로그 서열을 증폭시키도록 고안된 실시간 PCR 검정으로 수행될 수 있다 (문헌 [Deutsch, S. et al. 2004 J Med Genet 41, 908-915]). 이러한 맥락에서, 정보제공 웰은 염색체 21 또는 염색체 1 유전자좌에 대해 양성인, 그러나 둘 모두에 대해서 양성은 아닌 웰로서 정의된다. 정배수체 태아에 대해, 어느 하나의 유전자좌에 대해 양성인 정보제공 웰의 수는 대략 동등해야 한다. 삼염색체 21 태아에 대해, 염색체 1보다 염색체 21에 대해 양성인 웰의 과다표출이 있어야 한다. 과다표출의 정확한 비율은 하기 절에 기술되어 있다.

[0081] III. 태아 서열의 혼입률

[0082] 상기 기술된 방법 200 및 205의 실시양태의 단점은 태아 특이적 마커가 요구된다는 것이다. 따라서, 본 발명의 한 실시양태에서, 비-태아 특이적 마커가 사용된다. 비-태아 특이적 마커를 사용하기 위해서, 본 발명의 실시양태는 모체 혈장 (즉, 생물학적 샘플) 중 태아 DNA 분획 농도를 측정한다. 이러한 정보와 함께, 보다 유용한 P_r 값이 하기와 같이 계산될 수 있다.

[0083] 모체 혈장 중 태아 DNA의 낮은 분획 백분율로도, 삼염색체 21 태아는 모체 혈장으로 방출된 태아 DNA의 게놈-등가물 (GE) 당 염색체 21 서열의 추가적 양에 기여할 것이다. 예를 들어, 태아에 의해 기여된 5 GE/mL DNA와 50 GE/mL의 총 DNA (즉, 10% 태아 DNA 분획 농도)를 함유하는 정배수체 임신으로부터의 모체 혈장 샘플은 모체 혈장 (밀리리터) 당 염색체 21 서열의 총 100개의 카피 (90개의 모체 카피 + 10개의 태아 카피)를 함유해야 한다. 삼염색체 21 임신에 대해, 태아 GE는 각각 3개 카피의 염색체 21에 기여하여, 모체 혈장 중 염색체 21 서열의 총 105개의 카피/mL (90개의 모체 카피 + 15개의 태아 카피)가 되도록 한다. 따라서, 10% 태아 DNA 농도에서, 삼염색체 임신의 모체 혈장 중 염색체 21 유래 서열의 양은 정배수체 경우의 1.05배일 것이다. 따라서, 분석적 접근법이 이러한 적은 정도의 정량 차이를 결정하도록 개발될 수 있다면, 삼염색체 21 태아의 비침습성 출생 전 진단에 대한 다형성-독립적 시험이 가능할 것이다.

[0084] 따라서, 과다표출의 정도는 분석된 DNA 샘플 중 태아 DNA 분획 농도에 좌우될 것이다. 예를 들어, 태반 DNA를 분석하는 경우, 태아 게놈 중 이론적 RCD 비율은 3:2, 즉 1.5-배 차이이어야 한다. 그러나, 상기 기술된 것과 같이, 이론적 RCD 비율은 10% 태아 DNA를 함유하는 모체 혈장 샘플을 분석하는 경우 1.05로 감소할 것이다.

실험적으로 유도된 P_r은 단지 염색체 21 유전자좌에 양성인 웰의 수를 총 정보제공 웰의 수로 나누어 계산된다. 실험적으로 유도된 P_r은 계산된 P_r 및 이론적 RCD 비율과 함께 SPRT 분석에 적용된다.

[0085] 도 4는 본 발명의 실시양태에 따라 태아 핵산의 백분율을 사용하여 질환 상태를 결정하는 방법 400을 나타낸다. 단계 410에서, 태아 물질의 분획 백분율이 측정된다. 한 실시양태에서, 분획 백분율은 비-태아-특이적 마커 (즉, 모계 및 태아 모두에 존재하는 유전자 서열)와 관련하여 태아-특이적 마커 (예를 들어, Y-염색체, 유전자 다형성 마커 (예컨대, SNP), 태반 후성 기호 (placental epigenetic signatures))의 양을 측정하여 결정된다. 실시간 PCR, 디지탈 PCR, 서열분석 반응 (대규모 병렬 게놈 서열분석을 포함) 또는 임의의 다른 정량 방법에 의해 실질적 측정을 수행할 수 있다. 한 측면에서, 이 측정에 대해 잠재적으로 대립유전자 불균형이 있을 수 있는 유전자 표적을 사용하지 않는 것이 바람직하다.

[0086] 단계 420에서, 샘플을 희석시키고, 희석시킨 샘플을 웰에 넣고, 각 웰에서 반응을 측정하는 것을 비롯한 디지탈 PCR 또는 다른 측정 방법이 실시된다. 단계 430에서, PCR 결과는 상이한 참조 핵산 서열 (예컨대, 염색체 또는 대립유전자)의 마커를 식별하기 위해서 사용된다. 단계 440에서, 과다표출된 서열의 실제 비율 (P_r)이 계산된다. 단계 450에서, 질환 상태를 결정하기 위한 분별값은 샘플 중 태아 물질의 백분율을 사용하여 계산된다. 단계 460에서, 실제 P_r 및 분별값으로부터, 불균형이 존재하는지 여부가 결정된다.

[0087] 한 실시양태에서, 참조 핵산 서열의 분획 백분율은 디지탈 RNA-SNP 방법에 혼입된다. 따라서, 암 세포로 인한 LOH를 연구하는 경우, 50% 미만 암 세포를 갖는 종양 샘플로 실시할 수 있다. 이것은 또한 50% 초과 암 세포를 갖는 샘플에 대해서도 사용되어, 더욱 정확한 P_r을 수득함으로써 부정확한 진단으로 이끌 수 있는 잘못된 양성의 수를 감소시킬 수 있다. 또다른 실시양태에서, 태아 핵산 백분율은 디지탈 PCR 방법에 혼입되어, 태아가 모체 혈장 핵산 분석으로부터 부모의 유전자 돌연변이 (예를 들어, 낭포성 섬유증 또는 베타-지중해빈혈 또는 척수 근위축증을 유발하는 돌연변이) 또는 다형성을 물려받는지 여부를 결정한다.

[0089] IV. 웰 당 평균 농도의 혼입

[0090] 이전 방법 (예를 들어, 상기 문헌 [Zhou, W. et al. 2002])의 또다른 단점은 웰 당 주형의 평균 농도 (m)가 웰 당 1로 요구된다는 것이다. 정확한 농도를 얻는 것이 어렵다는 것을 감안할 때, 이것은 부정확해질 수 있다. 또한, 웰 당 1 주형의 정확한 농도로도, 이전 방법은 웰 중 주형의 통계적 분포를 간과하였다. 이전 방법에서, 즉, 기존 알고리즘에서, 대립가설을 수용하기 위한 예상 P_r 값은 대립유전자 비율로 가정되어, 웰 당 주형 DNA의 평균 농도와 무관하였다.

[0091] 그러나, 희석시킨 샘플에서 주형의 자연적인 통계적 편차로 인해, 정확하게 웰 당 1 주형은 아닐 것이다. 본 발명의 실시양태는 하나 이상의 서열의 평균 농도를 측정하여, 분별값, 즉 예상 P_r을 계산하는데 사용된다. 한 측면에서, 이러한 계산은 상이한 핵산 서열을 함유하는 웰의 확률을 측정하여 예상 P_r을 결정하기 위해서 사용되는 통계적 분포를 포함한다.

[0092] 한 실시양태에서, 평균 농도는 일례로 DNA 샘플 중 더 낮은 농도의 핵산 서열인 한 참조 핵산 서열로 고려된다. 불균형이 없는 샘플의 경우, 샘플 중 두 서열의 농도는 동일할 것이고, 어느 하나는 참조 대립유전자로 간주될 수 있다. 예를 들어 LOH가 있는 샘플의 경우, 암 세포에서 결실된 대립유전자는 참조 대립유전자로 간주될 것이다. 참조 대립유전자의 평균 농도는 m_r로 표시될 것이다. 또다른 실시양태에서, 더 높은 농도의 서열이 참조 서열로 간주될 수 있다.

[0094] **A. 디지탈-SNP, SPRT 및 디지탈 PCR을 사용한 예**

[0095] 도 5는 본 발명의 한 실시양태에 따라 평균 주형 농도를 사용하여 질환 상태를 결정하는 방법 500을 보여준다. 단계 510에서, 상이한 서열의 양을 측정한다. 이것은, 예를 들어 상기 설명된 것과 같은 디지탈 PCR 실험에서 마커를 계수하여 실시할 수 있다. 그러나, 이것은 증폭 단계를 포함하지 않거나 또는 형광 마커를 사용하지 않지만 다른 성질, 예컨대 질량, 특정 광학적 성질 또는 염기-쌍 성질과 같은 물성을 사용할 수 있는 다른 방법에 의해서도 실시할 수 있다.

[0096] 단계 520에서, 과다표출된 서열의 실제 비율이 결정된다. 이것은 상기 기술된 것과 같이 단지 그 서열만 나타내는 웰의 수를 취하고, 정보제공 웰의 수로 나누어 실시될 수 있다. 단계 530에서, 하나 이상의 서열의 평균 농도가 측정된다 (참조 서열). 한 실시양태에서, 참조 서열은 과다표출된 서열이다. 또다른 실시양태에서, 참

조 서열은 과소표출된(underrepresented) 서열이다. 측정은 디지탈 PCR 실험에서 참조 서열에 대해 음성인 웰의 수를 계수하여 실시할 수 있다. 음성 웰의 비율과 평균 주형 농도 사이의 관계는 다음 소절에 기술된 것과 같은 푸아송 (Poisson) 분포에 의해 설명된다.

[0097] 단계 540에서, 상이한 서열에 대해 양성인 웰의 예상 양이, 예를 들어 푸아송 분포를 사용하여 계산된다. 예상 양은 웰 당 서열, 웰 당 평균 서열, 서열을 함유하는 웰의 수 또는 임의의 다른 적합한 양의 확률일 수 있다. 단계 550에서, 예상 P_r은 예상 양으로부터 계산된다. 단계 560에서, 분별값은, 예를 들어 SPRT를 사용하여 예상 P_r로부터 계산된다. 단계 570에서, 핵산 서열 불균형의 분류가 결정된다. 방법 500의 특정 측면이 이제 기술된다.

[0099] 1. 서열 예상 양의 결정

[0100] 웰 (반응 또는 반응 혼합물) 당 평균 농도가 단계 530으로부터 알려지면, 서열을 나타내는 웰의 예상 수가 단계 540에서 계산될 수 있다. 이 양은 %, 분수 값, 또는 정수 값으로 표현될 수 있다. 설명을 위해 특정 예를 사용하면서, 웰 당 참조 주형의 평균 농도 (m_r)는 웰 당 0.5이고, PLAC4 SNP, rs8130833에서 삼염색체 21 태아의 유전자형은 AGG인 것으로 가정한다. 따라서, 참조 주형은 A 대립유전자일 것이고, 과다표출된 주형은 G 대립유전자일 것이다.

[0101] 한 실시양태에서, 푸아송 분포는 디지탈 PCR과 같은 측정 과정의 웰의 반응 혼합물 중 A 대립유전자의 분포로 가정한다. 다른 실시양태에서는, 이항 분포와 같은 다른 분포 함수가 사용된다.

[0102] 푸아송 방정식은 $P(n)=\frac{m^n e^{-m}}{n!}$ 이고, 여기서 n은 웰 당 주형 분자의 수이고; P(n)은 특정 웰 중 n 주형 분자의 확률이고; m은 특정 디지탈 PCR 실험에서 한 웰 중 주형 분자의 평균 수이다.

[0103] 따라서, 0.5의 평균 A-대립유전자 농도에서 A 대립유전자의 임의의 분자를 함유하지 않는 임의의 웰의 확률은 하기와 같을 것이다:

[0104] $$P(0)=\frac{0.5^0 e^{-0.5}}{0!}=e^{-0.5}\approx 0.6065$$

[0105] 따라서, 하나 이상의 A 대립유전자의 분자를 함유하는 임의의 웰의 확률은: 1 - 0.6065 = 0.3935이다. 그러므로, 약 39%의 웰이 하나 이상의 A 대립유전자의 분자를 함유할 것으로 예상될 것이다.

[0106] 비-참조 핵산 서열의 경우, 삼염색체 21 태아의 각 세포에 대해서, A 대 G의 게놈 비율은 1:2일 것이다. 추출된 RNA 또는 DNA 샘플 중 A 대 G 비율이 변화하지 않는 것으로 가정하면, 웰 당 G 대립유전자의 평균 농도는 A 대립유전자의 평균 농도의 2배, 즉 2 × 0.5 = 1일 것이다.

[0107] 따라서, 평균 G-대립유전자 농도인 1에서 G 대립유전자의 임의의 분자를 함유하지 않는 임의의 웰의 확률은 하기와 같을 것이다:

[0108] $$P(0)=\frac{1^0 e^{-1}}{0!}=e^{-1}\approx 0.3679$$

[0109] 따라서, 하나 이상의 G 대립유전자 분자를 함유하는 임의의 웰의 확률은: 1 - 0.3679 = 0.6321일 것이다. 그러므로, 약 63%의 웰이 하나 이상의 G 대립유전자 분자를 함유하는 것으로 예상될 것이다.

[0110] 2. 과다표출된 서열 비율의 결정

[0111] 예상 양이 계산된 후, 과다표출된 핵산 서열의 비율이 결정될 수 있다. A 대립유전자 및 G 대립유전자로 웰을 채우는 것이 독립적이라고 가정하면, 두 대립유전자를 모두 함유하는 웰의 확률은 0.3935 × 0.6321 = 0.2487일 것이다. 그러므로, 약 25%의 웰이 두 대립유전자를 모두 함유하는 것으로 예상될 것이다.

[0112] A 대립유전자는 함유하되 G 대립유전자는 함유하지 않는 것으로 예상되는 웰의 비율은 하나 이상의 A 대립유전자를 함유하는 웰의 수에서 A 및 G 대립유전자를 모두 함유하는 웰의 수를 뺀 것이다: 0.3935 - 0.2487 = 0.1448. 유사하게, G 대립유전자는 함유하되 A 대립유전자는 함유하지 않는 것으로 예상되는 웰의 비율은 0.6321 - 0.2487 = 0.3834일 것이다. 정보제공 웰은 A 대립유전자 또는 G 대립유전자에 대해 양성이나, 둘 모두에 대해서는 양성이 아닌 웰로서 정의된다.

[0113] 따라서, 디지탈 RNA-SNP 분석에서 G 대립유전자에 상대적으로 A 대립유전자를 함유하는 웰의 예상 비율은 0.1448/0.3834이다. 즉, 단지 G 대립유전자에 대해서만 양성인 웰의 비율은 단지 A 대립유전자에 대해서만 양성인 웰의 2.65배이다. 이것은 과다표출된 대립유전자가 다른 대립유전자의 2배인 태아 게놈 비율과 대조적이다.

[0114] SPRT 분석에 대해, 과다표출된 대립유전자에 대해 양성인 정보제공 웰의 비율 (P_r)은 SPRT 곡선을 사용하여 계산되고 해석된다. 현 예에서, 정보제공 웰의 비율은 0.1448 + 0.3834 = 0.5282일 것이다. 따라서, m_r 0.5에서 삼염색체 21 경우의 예상 P_r은 0.3834/0.5282 = 0.73이다.

[0115] 평균 주형 농도 (m)는 푸아송 방정식의 주요 파라미터이므로, P_r은 m에 따라변할 것이다. 도 6은 본 발명의 실시양태에 따라 웰 당 평균 참조 주형 농도 (m_r)로 표현된 주형 농도의 범위에 대한 삼염색체 21 샘플의 예상 디지탈 RNA-SNP 대립유전자 비율 및 P_r을 표로 만든 표 600을 나타낸다. 표 600은 일련의 웰 당 평균 참조 주형 농도 (m_r)에 대한 예상 대립유전자 비율 및 과다표출된 대립유전자에 대해 양성인 정보제공 웰의 비율 (P_r)을 나타낸다.

[0116] 예상 P_r 값은 비-선형 방식으로 웰 당 참조 대립유전자의 평균 농도 (m_r)에 따라 변한다. 표 600에 제시된 것과 같이, 대립가설을 수용하기 위한 예상 P_r 값은 m_r과 함께 증가할 것이다. 귀무가설을 수용하기 위한 예상 P_r 값이 0.5로 고정되어 있으므로, 대립유전자 불균형이 있는 샘플 및 대립유전자 불균형이 없는 샘플은 m_r이 증가할 경우, P_r 값의 측면에서 보다 광범위하게 나뉘어질 것이다. 나른 실시양태에서는 귀무가설을 수용하기 위한 값이 0.5가 아닐 수 있다는 것을 주목한다. 이것은 정상 비율이 1:1과 다른 경우, 예를 들어, 5:3인 경우 일어날 수 있으며, 따라서 비율이 5:3으로부터 벗어나는 경우 불균형이 발생할 것이다. 두 상이한 핵산 서열 양의 차이는 개별 경우에 따라 결정될 것이다.

[0117] 그러나, 이전 방법 (예를 들어, 상기 문헌 [Zhou, W. et al. 2002])은 LOH 샘플에 대해 고정된 예상 P_r 값을 사용하였으므로, LOH가 있는 샘플에 대한 P_r 값을 과소평가하였다 (대립가설 수용). 과소평가의 정도는 m_r과 함께 증가할 것이다. 즉, DNA 샘플 중 참조 대립유전자의 평균 농도가 더 높을수록, 이전 방법은 더욱 부정확해질 것이다. 대립가설을 수용하기 위한 P_r의 과소평가는 귀무가설 및 대립가설 모두를 수용하기 위한 분별값의 계산을 부정확하게 할 것이다.

[0119] 3. 예상 P_r을 기준으로 한 분별값의 계산

[0120] SPRT를 사용하는 실시양태에 대해, 문헌 [El Karoui at al. (2006)]으로부터 SPRT 곡선의 상한 및 하한을 계산하기 위한 방정식을 사용할 수 있다. 또한, 귀무가설 또는 대립가설을 수용하기 위해서 바람직한 통계적 신뢰성의 수준은 방정식에서 임계값 우도 비율(threshold likelihood ratio)을 조정하면서 변할 수 있다. 이 연구에서, 임계값 우도 비율 8이 암 검출의 맥락에서 대립유전자 불균형이 있는 샘플과 대립유전자 불균형이 없는 샘플을 구별하는 만족할만한 성능을 제공하는 것으로 나타났기 때문에, 상기 값 8이 사용된다. 따라서, 한 실시양태에서, SPRT 곡선의 상한 및 하한을 계산하기 위한 방정식은 다음과 같다:

[0121] 상한 = $\lceil (\ln 8)/N - \ln \delta \rceil / \ln \gamma$

[0122] 하한 = $\lceil (\ln 1/8)/N - \ln \delta \rceil / \ln \gamma$

[0123] 여기서, $\delta = (1 - \Theta_1)/(1 - \Theta_0)$,

[0124] $\gamma = -(\Theta_1(1 - \Theta_0)/\Theta_0(1 - \Theta_1)$,

[0125] Θ_0 = 귀무가설이 사실인 경우 비-참조 대립유전자를 함유하는 정보제공 웰의 비율 = 0.5 (하기 참조),

[0126] Θ_1 = 대립가설이 사실인 경우 비-참조 (즉, 과다표출된) 대립유전자를 함유하는 정보제공 웰의 비율,

[0127] N = 정보제공 웰의 수 = 어느 하나의 대립유전자에만 양성인 웰의 수.

[0128] (ln은 자연 로그, 즉 $\log_e$를 나타내는 수학 기호임)

[0129] 귀무가설을 수용하기 위한 Θ_0의 결정을 위해서, 샘플은 정배수체 태아를 갖는 임신 여성으로부터 일어지는 것으로 가정된다. 이러한 가정 하에, 어느 하나의 주형에 양성인 웰의 예상 수는 1:1일 것이므로, 비-참조 대립유전자를 함유하는 정보제공 웰의 예상 비율은 0.5일 것이다.

[0130] 대립가설을 수용하기 위한 Θ_1의 결정을 위해서, 샘플은 삼염색체 21 태아를 갖는 임신 여성으로부터 일어지는 것으로 가정된다. 디지탈 RNA-SNP 분석을 위한 삼염색체 21 경우의 예상 P_r의 계산은 표 600에 상술되어 있다. 따라서, 디지탈 RNA-SNP 분석을 위한 Θ_1은 표 600의 마지막 열에 제시된 데이타를 참조한다.

[0132] 4. 평균 농도의 측정

[0133] m_r의 측정은 당업자에게 공지되거나 또는 공지될 다양한 메커니즘에 의해 실시될 것이다. 한 실시양태에서, m_r 값은 디지탈 PCR 분석의 실험 과정 중 결정된다. m_r 값과 참조 대립유전자에 대해 양성인 웰의 총 수 사이의 관계는 분포 (예를 들어, 푸아송 분포)에 의해 좌우될 수 있으므로, m_r은 다음 식을 사용하여 참조 대립유전자에 대해 양성인 웰의 수로부터 계산될 수 있다:

[0134] m_r = - ln (1 - 참조 대립유전자에 대해 양성인 웰의 비율).

[0135] ln은 자연 로그, 즉 $\log_e$라는 것에 주목한다. 이 접근법은 디지탈 PCR 실험을 위해 사용되는 DNA 샘플 중 m_r의 직접적이고 정확한 평가를 제공한다.

[0136] 이 방법은 바람직한 농도를 얻기 위해서 사용될 수 있다. 예를 들어, 샘플의 추출된 핵산은 방법 200의 단계 240에서 실시한 것과 같이 반응 웰 당 하나의 주형 분자와 같은 특정 농도로 희석될 수 있다. 푸아송 분포를 사용하는 한 실시양태에서, 주형이 없는 웰의 예상 비율은 e^{-m}으로 계산될 수 있고, 여기서 m은 웰 당 주형 분자의 평균 농도이다. 예를 들어, 웰 당 하나의 주형 분자의 평균 농도에서, 주형 분자가 없는 웰의 예상 비율은 e^{-1}, 즉 0.37 (37%)로 주어진다. 나머지 63%의 웰은 하나 이상의 주형 분자를 함유할 것이다. 전형적으로, 디지탈 PCR 실행에서 양성 웰 및 정보제공 웰의 수가 계수될 것이다. 정보제공 웰의 정의 및 디지탈 PCR 데이타가 해석되는 방식은 용도에 좌우된다.

[0137] 다른 실시양태에서, 웰 당 평균 농도 m_r은 또다른 정량화 방법, 예를 들어 질량 분광법을 사용하는 정량 실시간 PCR, 반-정량 경쟁적 PCR, 실-경쟁적 PCR 등에 의해 측정된다.

[0139] **B. 디지탈 RCD**

[0140] 평균 농도를 사용하는 디지탈 RCD를 상기 기재된 디지탈 SNP 방법과 유사한 방식으로 수행할 수 있다. 참조 염색체 (비-염색체 21) 마커, 염색체 21 마커, 및 두 마커 모두에 대해 양성인 웰의 수를 디지탈 PCR에 의해 측정할 수 있다. 디지탈 SNP 분석에 대한 웰 당 참조 마커의 평균 농도 (m_r)의 계산에서와 같은 푸아송 확률 함수에 따라, 염색체 21 마커의 양성에 상관없이, 웰 당 참조 마커의 평균 농도 (m_r)를 참조 마커에 대해 음성인 웰의 총 수로부터 계산할 수 있다.

[0141] 이후, 정배수체 또는 삼염색체 21 태아를 갖는 임신 여성으로부터 얻어진 것과 같은 혈장 샘플을 분류하기 위해 SPRT 분석을 사용할 수 있다. 태아가 정배수체인 경우 귀무가설이 수용될 것이다. 이 시나리오에서, 참조 마커 및 염색체 21 마커에 대해 양성인 웰의 예상 비율은 1:1일 것이며, 이에 따라, 염색체 21 마커에 대해 양성 신호를 갖는 정보제공 웰의 예상 비율은 0.5일 것이다. 태아가 염색체 21에 대해 삼염색체인 경우 대립가설이 수용될 것이다. 이 시나리오에서, 샘플 DNA가 전적으로 태아로부터 유래된 경우, 각 웰에서의 염색체 21 마커의 평균 농도는 참조 마커의 평균 농도 (m_r)의 3/2배일 것이다.

[0142] 디지탈 RCD를 사용하여 태아-특이적 마커, 예를 들어 태반 후성 기호의 검출을 통해 염색체 양을 측정할 수 있지만 (문헌 [Chim, SSC. et al. 2005 Proc Natl Acad Sci USA 102, 14753-14758]), 디지탈 RCD 분석의 실시양태는 비-태아-특이적 마커를 사용한다. 따라서, 비-태아 특이적 마커를 사용하는 경우 태아 물질의 백분율을 측정하는 추가의 단계가 나타날 것이다. 따라서, 웰 당 염색체 21 마커의 평균 농도는 샘플 중 태아 DNA의 비율에 의존적일 것이며, m_r [(200% + 태아 DNA 백분율)/200%]을 사용하여 계산할 수 있다.

[0143] 예시용 특정 실시예를 다시 사용하여, 웰 당 참조 주형인 염색체 1의 평균 농도 (m_r)는 0.5인 것으로 가정하며,

50%의 DNA는 태아로부터 유래된 것으로 가정하고, 샘플 중 50%의 DNA는 모체로부터 유래된다.

[0144] 따라서, 푸아송 분포를 사용하여, 평균 농도가 웰 당 0.5인 경우 염색체 1 유전자좌의 어떠한 분자도 함유하지 않는 임의의 웰의 확률은 하기와 같을 것이다:

$$P(0)=\frac{0.5^{0}e^{-0.5}}{0!}=e^{-0.5}=0.6065$$

[0145]

[0146] 따라서, 염색체 1 유전자좌 중 하나 이상의 분자를 함유하는 임의의 웰의 확률은 1 - 0.6065 = 0.3935일 것이다. 따라서, 대략 39%의 웰이 유전자좌 중 하나 이상의 분자를 함유하는 것으로 예상될 것이다.

[0147] 이러한 삼염색체 21 태아의 각각의 세포에 대하여, 염색체 21 대 염색체 1의 게놈 비율은 3:2일 것이다. DNA 샘플에서 염색체 21과 염색체 1 사이의 비율은 태아 DNA 분획 농도 (태아 DNA %)에 의존적일 것이며, 3×태아 DNA % + 2(1 - 태아 DNA %) : 2×태아 DNA % + 2×(1 - 태아 DNA %)일 것이다. 따라서, 이 경우에서, 태아 DNA 분획 농도가 50%인 경우, 비는 (3×50% + 2×50%) / (2×50% + 2×50%) = 1.25일 것이다. 디지탈 SNP 방법이 태아 특이적 마커를 사용하지 않은 경우, 이러한 계산을 사용하여 비-참조 서열의 평균 농도를 계산할 수도 있다.

[0148] 따라서, 웰 당 염색체 1 유전자좌의 평균 농도가 0.5인 경우, 웰 당 염색체 21 유전자좌의 평균 농도는 1.25×0.5 = 0.625이다. 따라서, 평균 농도가 웰 당 0.625인 경우 염색체 21 유전자좌의 어떠한 분자도 함유하지 않는 임의의 웰의 확률은 하기와 같을 것이다:

$$P(0)=\frac{0.625^{0}e^{-0.625}}{0!}=e^{-0.625}=0.5353$$

[0149]

[0150] 따라서, 염색체 21 유전자좌 중 하나 이상의 분자를 함유하는 임의의 웰의 확률은 1 - 0.5353 = 0.4647일 것이다. 따라서, 대략 46%의 웰이 유전자좌 중 하나 이상의 분자를 함유하는 것으로 예상될 것이다. 유전자좌 중 하나로의 웰의 충전이 독립적이라고 가정하면, 유전자좌 모두를 함유하는 웰의 확률은 0.3935×0.4647 = 0.1829일 것이다. 따라서, 대략 18%의 웰이 두 유전자좌 모두를 함유하는 것으로 예상될 것이다.

[0151] 염색체 1 유전자좌를 함유하되 염색체 21 유전자좌는 함유하지 않는 것으로 예상되는 웰의 비율은, 하나 이상의 염색체 1 유전자좌를 함유하는 웰의 수에서 유전자좌 모두를 함유하는 웰의 수를 차감한 것, 즉 0.3935 - 0.1829 = 0.2106일 것이다. 유사하게, 염색체 21 유전자좌를 함유하되 유전자좌 모두는 함유하지 않는 것으로 예상되는 웰의 비율은 0.4647 - 0.1829 = 0.2818일 것이다. 정보제공 웰은 염색체 1 유전자좌 또는 염색체 21 유전자좌 중 하나에 대해 양성이되 둘 모두에 대해서는 아닌 웰로서 정의된다.

[0152] 따라서, 디지탈 RCD 분석에서의 염색체 21 대 염색체 1 예상 비율은 0.2818/0.2106 = 1.34이다. 바꾸어 말하면, 염색체 21 유전자좌에 대해서만 양성인 웰의 비율은 염색체 1 유전자좌에 대해서만 양성인 웰의 비율의 1.34배이다. 이는 DNA 샘플에서의 1.25의 비율과는 대조적이다.

[0153] SPRT 분석에 대하여, 염색체 21 유전자좌에 대해 양성인 정보제공 웰의 비율 (P_r)을 SPRT 곡선을 사용하여 계산 및 해석하는 것이 필요할 것이다. 현재 예에서, 정보제공 웰의 비율은 0.2106 + 0.2818 = 0.4924일 것이다. 따라서, 0.5의 m_r에서 50% 태아 DNA를 갖는 삼염색체 21 경우의 예상 P_r은 0.2818/0.4924 = 0.57이다.

[0154] 평균 주형 농도 (m)가 푸아송 방정식에서의 중요 파라미터이기 때문에, P_r은 m에 의해 변할 것이다. 도 7은 본 발명의 실시양태에 따른 웰 당 평균 참조 주형 농도 (m_r)로서 표현된 주형 농도의 범위에서 삼염색체 21 샘플 중 10%, 25%, 50% 및 100%의 태아 DNA 분획 농도에 대한 예상 P_r을 표로 만든 표 700을 나타낸다. 디지탈 RCD 분석에 대한 삼염색체 21 경우의 예상 P_r에 대한 계산을 표 700에서 상술한다. 따라서, 표 700에서 상응하는 예상 P_r 값을 나타내는 열로부터 어러 태아 DNA 분획 농도를 갖는 샘플의 디지탈 RCD 분석에 대한 Θ_1을 읽을 수 있다.

[0156] **C. 결과**

[0157] <u>1. 상이한 m_r의 비교</u>

[0158] 대립유전자 또는 염색체 불균형의 이론값 (태아 게놈에서와 같음)과 실험적으로 예상되는 정도 사이의 차이, 및 m_r 값 범위에 대한 후자의 측정을 위한 계산값에 대한 근거를 표 600 및 700에 나타낸다. 삼염색체 21 샘플의 디지탈 RNA-SNP 분석에서, m_r = 0.5인 경우, 참조 대립유전자만을 함유하는 웰에 대한 과다표출된 대립유전자만을 함유하는 웰, 즉 디지탈 RNA-SNP 비율은 2.65이다 (표 600). 100% 태아 DNA로 구성된 표본의 디지탈 RCD 분석에서, m_r = 0.5인 경우, 염색체 1 유전자좌에 대해 단독으로 양성인 웰에 대한 염색체 21 유전자좌에 대해 단독으로 양성인 웰, 즉 디지탈 RCD 비율은 1.7이다 (표 700) (P_r = 0.63, 따라서 디지탈 RCD 비율은 0.63/(1 - 0.63) = 1.7임). 태아 DNA 분획 농도가 감소하기 때문에, 디지탈 RCD 비율은 동일한 m_r에 대해 감소한다 (표 700).

[0159] 표 600 및 700에서 나타낸 것과 같이, 대립유전자 또는 염색체 과다표출 정도는 m_r에 따라 증가한다. 그러나, 정보제공 웰의 백분율은 m_r = 0.5 근처에서 그의 최대에 도달하며, m_r이 추가로 증가함에 따라 점차 감소한다. 실제로, 표본에 대한 주형 분자의 양이 제한되지 않는 경우 정보제공 웰의 비율의 감소는 분석된 웰의 총 수의 증가에 의해 상쇄될 수 있지만, 추가의 웰은 시약 비용의 증가를 요구할 것이다. 따라서, 최적 디지탈 PCR 수행은 주형 농도와 샘플 당 시험된 웰의 총 수 간에 상호 의존적이다.

[0161] 2. SPRT 곡선을 사용하는 예

[0162] 상기 논의된 것과 같이, 디지탈 PCR 실험에 대한 예상 대립유전자 또는 염색체 불균형 정도는 반응 혼합물 (예를 들어, 웰) 당 실제 주형 농도에 의존적이다. 본 발명자들은 참조 대립유전자를 기준으로 한 주형 농도, 즉 웰 당 평균 참조 주형 농도 (m_r)를 기재한다. 상기 방정식에서 나타낸 것과 같이, 예상 P_r을 사용하여 상부 및 하부 SPRT 곡선의 플롯팅을 결정할 수 있다. 예상 P_r이 또한 m_r 값에 의존적이기 때문에, SPRT 곡선의 플롯팅은 본질적으로 m_r 값에 의존적일 것이다. 따라서, 실제로, 디지탈 PCR 데이타 세트의 실제 m_r에 관련된 한 세트의 SPRT 곡선이 특정 실행으로부터 P_r의 해석을 위해 사용될 필요가 있다.

[0163] 도 8은 본 발명의 실시양태에 따른 디지탈 RNA-SNP 분석에 있어서 0.1, 0.5 및 1.0의 m_r 값에 대한 SPRT 곡선에서의 차이의 정도를 예시하는 플롯 800을 나타낸다. 각 세트의 디지탈 PCR 데이타는 특정 실행의 정확한 m_r 값에 관련된 특정 곡선으로 해석되어야 한다. 디지탈 RNA-SNP 및 RCD 방법에 대한 예상 대립유전자 또는 염색체 불균형 정도가 상이하기 때문에 (전자: 2:1, 후자: 3:2), 2개 디지탈 PCR 시스템에 대하여 상이한 세트의 SPRT 곡선이 필요하다는 것을 주목한다. 실험적으로 유도된 P_r을 디지탈 PCR 실행의 상응하는 m_r에 의해 선택된 관련 SPRT 곡선을 이용하여 해석한다. 이는 고정된 세트의 곡선이 사용된 경우 디지탈 PCR에 의한 LOH의 분자 검출을 위해 앞서 보고된 SPRT를 사용한 것과는 대조적이다.

[0164] SPRT를 사용하여 디지탈 PCR 데이타를 해석하기 위한 실제적 방법을 가상 디지탈 RNA-SNP 실행을 사용하여 하기 예시한다. 각 경우의 디지탈 RNA-SNP 분석 이후, A 대립유전자에 대해서만 양성인 웰, G 대립유전자에 대해서만 양성인 웰, 또는 두 대립유전자 모두에 대해서 양성인 웰의 수를 계수한다. 참조 대립유전자는 보다 적은 수의 양성 웰을 갖는 대립유전자로서 정의된다. 푸아송 확률 밀도 함수에 따라, 다른 대립유전자가 양성인 지에 관계없이, 참조 대립유전자에 대해 음성인 웰의 총 수를 사용하여 m_r 값을 계산한다. 본 발명자의 가설 예의 데이타는 다음과 같다:

[0165] 96-웰 반응에서, 20개의 웰은 A 대립유전자에 대해서만 양성이고, 24개의 웰은 G 대립유전자에 대해서만 양성이며, 33개의 웰은 두 대립유전자 모두에 대해 양성이다. A 대립유전자가 참조 대립유전자로 간주되는데, A-양성 웰이 G-양성 웰에 비해 적기 때문이다. 참조 대립유전자에 대해 음성인 웰의 수는 96 - 20 - 33 = 43이다. 따라서, 푸아송 방정식을 사용하여 다음과 같이 m_r을 계산할 수 있다: $-\ln(43/96) = 0.80$. 이 경우의 실험적으로 결정된 P_r은 24/(20 + 24) = 0.55이다.

[0166] 표 600에 따르면, m_r = 0.8에서의 삼염색체 21 샘플의 예상 P_r은 0.76이다. 따라서, 이 경우 Θ_1은 0.76이다. O_1 = 0.76을 기준으로 한 SPRT 곡선을 사용하여 실험적으로 유도된 P_r (이 경우 0.55)을 해석할 것이다. P_r = 0.55를 관련된 SPRT 곡선 상에 피팅하는 경우, 데이타 지점은 하부 곡선 아래에 있다. 따라서, 본 경우는 정배수체로 분류되며, 도 3을 참조한다.

[0168] 3. 기존 방법과의 비교

[0169] 도 9a는 96-웰 디지탈 RNA-SNP 분석에서의 정배수체 및 삼염색체 21 경우를 분류하기 위한 신규 및 기존 SPRT 알고리즘의 효과의 비교 표 900을 나타낸다. 도 9b는 384-웰 디지탈 RNA-SNP 분석에서 정배수체 및 삼염색체 21 경우를 분류하기 위한 신규 및 기존 SPRT 알고리즘의 효과의 비교의 표 950이다. 신규 알고리즘은 디지탈 PCR 데이타로부터 유도된 m_r에 특이적인 SPRT 곡선의 선택을 의미한다. 기존 알고리즘은 모든 디지탈 PCR 실행에 대한 고정된 세트의 SPRT 곡선의 사용을 의미한다. 분류의 정확도에 대한 분별값의 부정확한 계산의 효과는 표 900에 나타낸 시뮬레이션 분석에 의해 밝혀진다.

[0170] 기존 연구에서 고정된 세트의 SPRT 곡선의 사용과 비교하여, 분류불가능한 데이타의 비율은 표 900 및 950에서 나타난 것과 같이 본 발명의 방법에서 훨씬 적다. 예를 들어, 본 발명의 방법을 사용하여, m_r = 0.5에서, 14% 및 0%의 삼염색체 21 샘플이 각각 96-웰 및 384-웰 디지탈 RNA-SNP 분석에 대하여 분류불가능한 것이나, 고정된 곡선의 사용에 대해서는 각각 62% 및 10%가 분류불가능할 것이다 (표 900). 따라서, 본 발명의 방법은 보다 적은 수의 정보제공 웰로 질환 분류를 가능하게 한다.

[0171] 표 900에서 나타낸 것과 같이, 신규 알고리즘은 0.1 내지 2.0의 모든 m_r 값에 대한 대립유전자 비율이 벗어나거나 벗어나지 않은 샘플의 분류에서 보다 정확하다. 예를 들어, m_r이 1.0과 같고, 96-웰 디지탈 RNA-SNP 실행을 수행하는 경우, 신규 알고리즘은 각각 88% 및 92%의, 대립유전자 비율이 벗어난 샘플 및 벗어나지 않은 샘플을 정확하게 분류하며, 반면 기존 알고리즘을 사용하여 대립유전자 비율이 벗어난 샘플 및 벗어나지 않은 샘플의 정확한 분류의 백분율은 단지 각각 19% 및 36%이다.

[0172] 신규 알고리즘을 사용하여, 대립유전자 비율이 벗어난 샘플과 벗어나지 않은 샘플의 분리는 m_r에 따라 증가할 것이다. 결과적으로, 분류 정확도는 m_r에 따라 증가할 것이다. m_r이 2.0 초과로 증가하는 경우 분류 정확도에 대한 2개 샘플 그룹의 분리에서의 증가의 효과는 감소될 것인데, 이는 정보제공 웰의 백분율에서의 감소 때문이다. 반면, 기존 알고리즘의 사용에서, m_r이 증가하는 경우 분류 정확도는 유의하게 감소하는데, 이는 예상 P_r 값과 그의 참값의 편차 증가 때문이다.

[0173] 본 발명자의 실험 및 시뮬레이션 데이타는 디지탈 RNA-SNP가 삼염색체 21 검출에 대해 효과적이며 정확한 방법임을 나타낸다. 모체 혈장에서의 *PLAC4* mRNA가 태아로부터 순수하게 유래되기 때문에, 시험된 13개 모체 혈장 샘플 중 12개에 대하여, 오직 1회의 384-웰 디지탈 PCR 실험만이 정확한 분류에 필요하였다 (도 13b의 표 1350). 이에 따라, 상기 균질한 실시간 디지탈 PCR-기반 방법은 RNA-SNP 분석에 대한 질량 분광법-기반 방법의 또다른 형태를 제공한다 (상기 문헌 [Lo, YMD, et al. 2007 Nat Med]). 태반-특이적 mRNA 전사체와는 별개로, 본 발명자들은 또한 모체 혈장에서의 다른 유형의 태아-특이적 핵산 종을 태아 염색체 이수성의 디지탈 PCR-기반 검출에 사용할 수 있다는 것을 구상한다. 한 예는 태아 후성 마커이며 (문헌 [Chim, SSC et al. (2005) Proc Natl Acad Sci USA 102, 14753-14758]; [Chan, KCA et al. (2006) Clin Chem 52, 2211-2218]), 이는 최근에 후성 대립유전자 비율 (EAR) 방법을 사용하여 삼염색체 18의 비-침습적 출생 전 검출에 사용되었다 (문헌 [Tong, YK et al. (2006) Clin Chem 52, 2194-2202]). 따라서, 본 발명자들은 디지탈 EAR이 가능한 분석 기법일 것으로 예상한다.

[0175] V. 증가 %, 다중 마커 및 PCR 변형법

[0176] 상기 기재된 것과 같이, 태아 DNA만이 임신 11주 내지 17주 사이에 대략 3%의 평균 분획 농도로 모체 혈장 DNA의 소수 분획을 구성하는 경우, 모체 혈장으로부터 추출된 DNA에 본 발명의 실시양태를 적용하는 것이 복잡해질 수 있다. 그러나, 본원에서 나타낸 것과 같이, 디지탈 RCD는 심지어 이수체 DNA가 소수 집단으로서 존재하는 경우에도 이수성 검출을 가능하게 한다. 임신 초기 동안 존재할 수 있는, 태아 DNA 분획 농도의 감소에 대하여, 보다 많은 수의 정보제공 계수가 디지탈 RCD에 대해 요구된다. 도 12의 표 1200에서 요약한 것과 같이, 본 작업의 의미는 본 발명자들이 한 세트의 기준 파라미터, 예를 들어 요구되는 태아 DNA 분획 및 총 주형 분자를 제공하였다는 것이며, 이들에 대해 진단 검정이 수립될 수 있다. 본 발명자의 견해에서, 25%의 태아 DNA 분획 농도에 대한 7680개의 총 수의 반응은 특히 매력적인 세트의 기준 파라미터이다. 이러한 파라미터는 표 1200에서 나타낸 것과 같이 정배수체 및 삼염색체 21 샘플이 정확하게 97%로 분류가능하게 한다.

[0177] 모체 혈장의 단위 부피 당 존재하는 혈장 DNA 분자의 수는 제한된다 (문헌 [Lo, YMD. et al. 1998 Am J Hum Genet 62, 768-7758]). 예를 들어, 임신 초기에서, 태아 및 모체 모두로부터의 기여에 대하여, 상염색체 유전

자좌, β-글로빈 유전자의 중앙 모체 혈장 농도는 986 카피/mL인 것으로 나타났다 (문헌 [Lo, YMD. et al. 1998 Am J Hum Genet 62, 768-7758]). 7,680개의 분자를 포획하기 위해, 대략 8 mL의 모체 혈장으로부터 추출된 DNA가 필요할 것이다. 약 15 mL의 모체 혈액으로부터 얻을 수 있는 상기 혈장 부피는 일반적인 실시의 한도이다. 그러나, 본 발명자들은 다중 세트의 chr21 및 참조 염색체 표적을 디지탈 RCD 분석을 위해 합할 수 있다는 것을 구상한다. 5쌍의 chr21 및 참조 염색체 표적에 대하여, 단지 1.6 mL의 모체 혈장이 분석에 필요한 주형 분자의 수를 제공하기 위해 필요할 것이다. 밀티플렉스 단일 분자 PCR을 수행할 수 있다. 이러한 밀티플렉스 단일 분자 분석의 견고성(robustness)은 단일 분자 일배체형화(haplotyping)에 대해 이미 입증되었다 (문헌 [Ding, C. and Cantor, CR. 2003 Proc Natl Acad Sci USA 100, 7449-7453]).

[0178] 별법으로, 25%의 태아 DNA 분획 농도를 달성하기 위해, 방법은 모체 혈장에서 태아 DNA의 선택적 풍부화 (문헌 [Li, Y. et al. 2004 Clin Chem 50, 1002-1011]) 또는 모체 DNA 배경의 억제 (문헌 [Dhallan, R et al. 2004 JAMA 291, 1114-1119]) 또는 둘 모두를 가능하게 할 수 있다. 태아 DNA 풍부화 및 모체 DNA 억제를 위한 물리적 방법과는 별개로, 특정 DNA 메틸화 패턴을 나타내는 태아 DNA 분사의 표적화와 같은 분사 풍부화 전략을 사용하는 것이 또한 가능할 것이다 (문헌 [Chim, SSC et al. 2005 Proc Natl Acad Sci USA 102,14753-14758], [Chan, KCA et al. 2006 Clin Chem 52, 2211-2218]; [Chiu, RWK et al. 2007 Am J Pathol 170, 941-950.])

[0179] 추가적으로, 디지탈 PCR 수행에 대한 현행 연구에 사용되는 것과 같은 디지탈 실시간 PCR 분석의 수동 설정을 위한 다수의 또다른 방법이 현재 존재한다. 이러한 다른 방법에는 미세유체 디지탈 PCR 칩 (문헌 [Warren, L et al. 2006 Proc Natl Acad Sci USA 103, 17807-17812]; [Ottesen, EA et al. 2006 Science 314, 1464-1467]), 에멀젼 PCR (문헌 [Dressman, D et al. 2003 Proc Natl Acad Sci USA 100, 8817-8822]), 및 예를 들어 로슈(Roche) 454 플랫폼, 일루미나 솔렉사(Illumina Solexa) 플랫폼, 및 어플라이드 바이오시스템즈(Applied Biosystems)의 솔리드(SOLiD)™ 시스템을 사용하는 대규모 병렬 게놈 서열분석 (문헌 [Margulies, M. et al. 2005 Nature 437, 376-380]) 등이 있다. 후자에 대하여, 본 발명자의 방법은 또한 단일 DNA 분자에 대한 대규모 병렬 서열분석 방법에 적용가능하며, 이는 증폭 단계, 예컨대 헬리코스 트루(Helicos True) 단일 분자 DNA 서열분석 기술 (문헌 [Harris TD et al. 2008 Science, 320, 106-109]), 퍼시픽 바이오사이언시스(Pacific Biosciences)의 단일 분자, 실시간 (SMRT™) 기술, 및 나노공극(nanopore) 서열분석 (문헌 [Soni GV and Meller A. 2007 Clin Chem 53, 1996-2001])을 요구하지 않는다. 이들 방법을 사용하여, 디지탈 RNA-SNP 및 디지탈 RCD를 다수의 샘플에 대하여 빠르게 수행함으로써, 비-침습적 출생 전 진단에 대해 본원에서 제안된 방법의 임상적 실행가능성을 향상시킬 수 있다.

도면의 간단한 설명

[0181] 도 1은 디지탈 PCR 실험을 설명하는 흐름도이다.

도 2a는 본 발명의 실시양태에 따른 디지탈 RNA-SNP 및 RCD 방법을 예시한다.

도 2b는 암에서 자주 검출되는 염색체 이상의 예에 대한 표를 도시한다.

도 3은 본 발명의 실시양태에 따른 다운 증후군을 결정하는데 사용된 SPRT 곡선을 갖는 그레프를 예시한다.

도 4는 본 발명의 실시양태에 따른 태아 세포의 백분율을 사용하여 질환 상태를 결정하는 방법을 도시한다.

도 5는 본 발명의 실시양태에 따른 평균 농도를 사용하여 질환 상태를 결정하는 방법을 도시한다.

도 6은 본 발명의 실시양태에 따른 웰 당 평균 참조 주형 농도 (m_r)로서 표현된 주형 농도 범위에 대해 삼염색체 21 샘플의 예상되는 디지탈 RNA-SNP 대립유전자 비율 및 P_r을 표로 만든 표를 도시한다.

도 7은 본 발명의 실시양태에 따른 웰 당 평균 참조 주형 농도 (m_r)로서 표현된 주형 농도 범위에서 삼염색체 21 샘플 중 분획 태아 DNA 농도 10%, 25%, 50% 및 100%에 대해 예상되는 P_r을 표로 만든 표를 도시한다.

도 8은 본 발명의 실시양태에 따른 디지탈 RNA-SNP 분석에서 m_r 값 0.1, 0.5 및 1.0에 대해 SPRT 곡선의 차이 정도를 예시하는 플롯을 도시한다.

도 9a는 본 발명의 실시양태에 따른 96-웰 디지탈 RNA-SNP 분석에서 정배수체 및 삼염색체 21 사례를 분류하기 위한 신규/기존 SPRT 알고리즘 효과의 비교에 대한 표를 도시한다.

도 9b는 본 발명의 실시양태에 따른 384-웰 디지탈 RNA-SNP 분석에서 정배수체 및 삼염색체 21 사례를 분류하기

위한 신규/기존 SPRT 알고리즘 효과의 비교에 대한 표를 도시한다.

도 10은 본 발명의 실시양태에 따른 소정의 정보제공 카운트에 대해 정배수체 또는 이수체로서 정확하게 및 부정확하게 분류된 태아 및 분류가능하지 않은 태아의 백분율을 도시하는 표이다.

도 11은 본 발명의 실시양태에 따른 순수 (100%) 태아 DNA 샘플에 대한 디지탈 RCD 분석용 컴퓨터 시뮬레이션을 도시하는 표 1100이다.

도 12는 본 발명의 실시양태에 따른 상이한 분획 농도의 태아 DNA를 갖는 정배수체 또는 삼염색체 21 태아로부터의 샘플을 분류하기 위해 m_F=0.5에서 디지탈 RCD 분석의 정확성을 컴퓨터 시뮬레이션한 결과를 도시하는 표 1200이다.

도 13a는 본 발명의 실시양태에 따른 정배수체 및 삼염색체 21 임신 태반 조직에서의 디지탈 RNA-SNP 분석에 대한 표 1300을 도시한다.

도 13b는 본 발명의 실시양태에 따른 정배수체 및 삼염색체 21 임신으로부터의 모체 혈장의 디지탈 RNA-SNP 분석에 대한 표 1350을 도시한다.

도 14a 내지 14c는 본 발명의 실시양태에 따른 RCD 분석으로부터 생성된 분별값 곡선을 예시하는 플롯을 도시한다.

도 15a는 본 발명의 실시양태에 따른 정배수체 및 삼염색체 21 임신 태반 조직에서의 디지탈 RNA-SNP 분석에 대한 표를 도시한다.

도 15b는 본 발명의 실시양태에 따른 한 모체 혈장 샘플로부터의 12개 반응 패널의 디지탈 RNA-SNP 데이타에 대한 표를 도시한다.

도 15c는 본 발명의 실시양태에 따른 정배수체 및 삼염색체 21 임신으로부터의 모체 혈장의 디지탈 RNA-SNP 분석에 대한 표를 도시한다.

도 16a는 본 발명의 실시양태에 따른 정배수체 및 삼염색체 18 태반의 디지탈 RNA-SNP 분석에 대한 표를 도시한다.

도 16b는 본 발명의 실시양태에 따른 정배수체 및 삼염색체 18 태반에 대한 디지탈 RNA-SNP 데이타의 SPRT 해석을 도시한다.

도 17은 본 발명의 실시양태에 따른 정배수체 및 삼염색체 21 임신의 50% 태반/모체 혈액 세포 DNA 혼합물의 디지탈 RCD 분석에 대한 표를 도시한다.

도 18은 본 발명의 실시양태에 따른 정확한 분류에 대한 판정 경계값을 예시하는 SPRT 곡선을 도시한다.

도 19는 본 발명의 실시양태에 따른 정배수체 및 삼염색체 21 임신으로부터의 양수 샘플의 디지탈 RCD 분석에 대한 표를 도시한다.

도 20은 본 발명의 실시양태에 따른 정배수체 및 삼염색체 18 임신 (E = 정배수체; T18 = 삼염색체 18)으로부터의 태반 DNA 샘플의 디지탈 RCD 분석에 대한 표를 도시한다.

도 21은 본 발명의 실시양태에 따른 정배수체 및 삼염색체 21 임신 (E = 정배수체; T21 = 삼염색체 21; U = 미분류)의 50% 태반/모체 혈액 세포 DNA 혼합물의 멀티플렉스 디지탈 RCD 분석에 대한 표를 도시한다.

도 22a 및 22b는 본 발명의 실시양태에 따른 50% 정배수체 또는 삼염색체 21 태반 게놈 DNA/50% 모체 연막(buffy coat) DNA 혼합물의 멀티플렉스 디지탈 RCD 분석에 대한 표를 도시한다. 미분류(Unclass)는 분류가능하지 않음을 나타내고, T21은 삼염색체 21을 나타낸다.

도 23은 남성 및 여성 파트너 둘 다가 동일한 돌연변이를 전달하는 시나리오를 도시한다.

도 24a는 본 발명의 실시양태에 따른 여성/남성 및 남성/남성 DNA 혼합물의 디지탈 RMD 분석에 대한 표를 도시한다.

도 24b는 본 발명의 실시양태에 따른 25% 여성 및 75% 남성 DNA를 갖는 혼합물의 디지탈 RMD 분석에 대한 표를 도시한다.

도 25는 본 발명의 실시양태에 따른 HbE 돌연변이에 대해 모체 혈장 샘플을 모방하는 15% 내지 50% DNA 혼합

물의 디지탈 RMD 분석에 대한 표를 도시한다.

도 26a는 본 발명의 실시양태에 따른 CD41/42 돌연변이에 대해 모체 혈장 샘플을 모방하는 5% 내지 50% DNA 혼합물의 디지탈 RMD 분석에 대한 표를 도시한다.

도 26b는 본 발명의 실시양태에 따른 CD41/42 돌연변이에 대해 모체 혈장 샘플을 모방하는 20% DNA 혼합물의 디지탈 RMD 분석에 대한 표를 도시한다.

도 27은 본 발명의 실시양태에 따른 시스템 및 방법으로 이용가능한 예시적 컴퓨터 장치의 블록 다이아그램을 도시한다.

발명을 실시하기 위한 구체적인 내용

[0182] **실시예**

[0183] 하기 실시예는 청구된 발명을 예시하되 제한하지 않으려고 한다.

[0185] I. 컴퓨터 시뮬레이션

[0186] SPRT 방법을 사용하여 심염색체 21 진단의 정확도를 평가하기 위해 컴퓨터 시뮬레이션을 수행하였다. 마이크로소프트 엑셀 2003 소프트웨어 (마이크로소프트 코포레이션(Microsoft Corp.; 미국)) 및 윈도우용 SAS 9.1 소프트웨어 (에스에이에스 인스티튜트 인크.(SAS Institute Inc., 미국 노스 캐롤라이나주))를 사용하여 컴퓨터 시뮬레이션을 수행하였다. 디지탈 PCR의 수행은 참조 주형 농도 (m_r), 정보제공 계수의 수, 및 대립유전자 또는 염색체 불균형의 추정 정도 (P_r) 사이의 상호작용이다. 이들 변수 각각의 범위에 대하여 별도의 시뮬레이션을 수행하였다. 디지탈 RNA-SNP 및 디지탈 RCD에 대한 SPRT 곡선의 결성 경계가 상이하기 때문에, 2개 시스템에 대한 시뮬레이션 분석을 별도로 수행하였다.

[0187] 시뮬레이션된 각각의 디지탈 PCR 조건 (즉, m_r, 태아 DNA 분획 농도, 총 웰 수)에 대하여, 2회의 시뮬레이션을 수행하였다. 제1회는 시험된 샘플을 정배수체 태아를 갖는 임신 여성으로부터 얻은 시나리오를 시뮬레이션하였다. 제2회는 시험된 샘플을 삼염색체 21 태아를 갖는 임신 여성으로부터 얻은 경우의 시나리오를 시뮬레이션하였다. 각각의 시뮬레이션에서, 5000명의 태아의 시험을 시뮬레이션하였다.

[0189] A. RNA-SNP

[0190] 디지탈 RNA-SNP에 대하여, m_r = 0.1 내지 m_r = 2.0으로 384-웰 실험의 시뮬레이션을 수행하였다. 각각의 m_r 값에서, 본 발명자들은 5000명의 정배수체 및 5000명의 삼염색체 21 태아를 시험하는 시나리오를 시뮬레이션하였다. 주어진 m_r에 적절한 SPRT 곡선을 사용하여 10,000명의 태아를 분류하였다. 도 10은 본 발명의 실시양태에 따른, 정배수체 또는 이수체로서 정확하게 및 부정확하게 분류된 태아, 및 주어진 정보제공 계수에 대해 분류가능하지 않은 태아의 백분율을 나타내는 표 1000이다. 정배수체 및 이수체 경우의 진단에 대한 정확도는 0.5 내지 2.0 사이의 m_r에 대해 모두 100%이다. m_r = 0.1인 경우, 57% 및 88%의 정배수체 및 삼염색체 21 태아만이 384-웰의 분석 이후 정확하게 분류될 수 있었다.

[0191] 하기 단계에 기재된 것과 같이 시뮬레이션 데이타를 생성하였다.

[0192] 단계 1에서, 각 웰에 대하여, SAS 프로그램의 무작위 (푸아송) 함수 (www.sas.com/technologies/analytics/statistics/index.html)를 사용하여 2개 난수를 생성하여 각각 A 및 G 대립유전자를 나타내었다. 무작위 (푸아송) 함수는 0으로부터 출발하는 양의 정수 (즉, 0, 1, 2, 3,....)를 생성할 것이며, 생성되는 각 정수의 확률은 웰 당 대립유전자의 평균 농도를 나타내는 주어진 평균값에 대한 푸아송 확률 밀도 함수에 따른 상기 수의 확률에 의존하였다. A 대립유전자를 나타내는 난수가 0보다 큰 경우, 즉 A 대립유전자의 1개 이상의 분자를 함유한 경우, 웰은 A 대립유전자에 대해 양성으로서 간주되었다. 유사하게, G 대립유전자를 나타내는 난수가 0보다 큰 경우, 웰은 G 대립유전자에 대해 양성으로서 간주되었다.

[0193] 정배수체 태아를 갖는 임신 여성의 시나리오를 시뮬레이션하기 위해, A 대립유전자 및 G 대립유전자에 대한 난수를 생성하는데 동일한 평균 값을 사용하였다. 예를 들어, m_r = 0.5에서의 디지탈 RNA-SNP 분석을 시뮬레이션하는 분석에서, A 대립유전자 또는 G 대립유전자 중 하나에 대한 평균 값을 웰 당 0.5 분자의 대립유전자 중 하나에 대한 평균 농도를 의미하는 0.5와 동일하게 설정하였다. 푸아송 방정식을 사용하여, 0.5의 평균

농도에서, A 또는 G 대립유전자에 대해 양성인 웰의 비율은 동일할 것이며, 0.3935였다 (표 600 참조).

[0194] m_r = 0.5에서 삼염색체 21 태아를 갖는 임신 여성의 디지탈 RNA-SNP 분석을 시뮬레이션한 경우, 웰 당 과다표출된 대립유전자의 평균 농도는 참조 대립유전자의 평균 농도의 2배, 즉 1이 되는 것으로 예상될 것이다. 이 상황에서, 과다표출된 대립유전자에 대해 양성인 웰의 확률은 0.6321이었다 (표 600 참조).

[0195] 디지탈 PCR 웰에 대한 난수의 생성 이후, 웰을 하기 상태 중 하나로서 분류할 수 있었다:

[0196] a. A 및 G 대립유전자 모두에 대해 음성

[0197] b. A 및 G 대립유전자 모두에 대해 양성

[0198] c. A 대립유전자에 대해 양성이되, G 대립유전자에 대해 음성

[0199] d. G 대립유전자에 대해 양성이되, A 대립유전자에 대해 음성.

[0200] 단계 2에서, 원하는 수의 웰 (현재 시뮬레이션에 대하여 384-웰)을 생성할 때까지 단계 1을 반복하였다. A 대립유전자에 대해서만 양성인 웰의 수 및 G 대립유전자에 대해서만 양성인 웰의 수를 계수하였다. 덜 양성인 웰을 갖는 대립유전자는 참조 대립유전자로서 간주되며, 보다 양성인 웰을 갖는 대립유전자는 잠재적으로 과다표출된 대립유전자로서 간주되었다. 정보제공 웰의 수는 대립유전자 중 하나에 대해 양성이되, 모두에 대해서는 아닌 웰의 총 수였다. 이후, 잠재적으로 과다표출된 대립유전자를 함유하는 정보제공 웰의 비율 (P_r)을 계산하였다. 귀무가설 또는 대립가설을 수용하기 위한 관련 SPRT 곡선에 대한 상한 및 하한을 본 발명의 실시양태에 따라 계산하였다.

[0201] 단계 3에서, 정배수체 또는 삼염색체 21 태아를 갖는 임신 여성의 각각의 2개 시나리오에 대해 5000회 시뮬레이션을 수행하였다. 각각의 시뮬레이션은 임신 여성으로부터 얻어진 독립적인 생물학적 샘플로서 간주될 수 있었다. 표 1000에서, 정배수체 경우의 정확한 분류는 귀무가설이 수용되는 정배수체 경우를 의미하며, 정배수체 경우의 부정확한 분류는 대립가설이 수용되는 정배수체 경우를 의미한다. 유사하게, 대립가설이 수용되는 삼염색체 21 경우는 정확하게 분류된 것으로 간주되며, 귀무가설이 수용되는 삼염색체 21 경우는 부정확하게 분류된 것으로 간주된다. 두 군 모두에 대하여, 미리-명시된 웰의 총 수를 시뮬레이션한 이후 귀무가설 또는 대립가설 모두가 수용되지 않는 경우에, 미분류된 것으로 간주되었다.

[0202] 단계 4에서, 단계 1 내지 3을 0.1 내지 2.0 범위의 m_r에 대해 0.1 증가량으로 수행하였다.

[0204] **B. RCD**

[0205] 도 11은 본 발명의 실시양태에 따라 0.1 내지 2.0 범위의 m_r에서 순수한 (100%) 태아 DNA 샘플에 대한 디지탈 RCD 분석용 컴퓨터 시뮬레이션을 나타내는 표 1100이다. 태아 DNA 분획 농도가 낮아졌기 때문에, 염색체 21 과다표출 정도는 감소되며, 이에 따라 정확한 질환 분류를 위해 보다 많은 수의 정보제공 웰이 요구된다. 따라서, m_r = 0.5에서 384개 내지 7680개 웰 범위의 총 웰 수에 대하여 50%, 25% 및 10%의 태아 DNA 농도에 대해 추가로 시뮬레이션을 수행하였다.

[0206] 도 12는 본 발명의 실시양태에 따른 상이한 분획 농도의 태아 DNA를 갖는 정배수체 또는 삼염색체 21 태아로부터의 샘플의 분류에 대하여 m_r = 0.5에서 디지탈 RCD 분석의 정확도의 컴퓨터 시뮬레이션 결과를 나타내는 표 1200이다. 디지탈 RCD의 수행은 보다 높은 태아 DNA 분획 농도를 갖는 경우에 대해 보다 양호하다. 25% 태아 DNA 농도 및 7680개 총 수의 PCR 분석에서, 97%의 정배수체 및 이수체 경우 모두가 부정확한 분류 없이 분류가능할 것이다. 나머지 3%의 경우는 분류가 달성될 수 있을 때까지 추가 분석을 요구한다.

[0207] 디지탈 RCD 분석을 시뮬레이션하기 위한 절차는 디지탈 RNA-SNP 분석에 대해 기재된 것과 유사하였다. 시뮬레이션을 위한 단계를 하기 기재한다.

[0208] 단계 1에서, 푸아송 확률 밀도 함수 하의 2개의 난수를 생성하여 참조 유전자좌, 염색체 1 및 염색체 21 유전자좌를 나타내었다. 정배수체 태아를 갖는 대상체에 대하여, 염색체 1 및 염색체 21 유전자좌 모두의 평균 농도는 동일하였다. 상기 시뮬레이션 분석에서, 웰 당 각 유전자좌에 대한 0.5의 평균 주형 농도를 사용하였다. 삼염색체 21 태아를 갖는 대상체에 대하여, 상기 시뮬레이션에서의 m_r은 0.5였으나, 웰 당 염색체 21 유전자좌의 평균 농도는 표 700에서 나타낸 것과 같이 시험된 샘플에서의 태아 DNA 분획 농도에 의존할 것이다. 웰에 대한 참조 및/또는 염색체 21 유전자좌의 분포를 웰 당 유전자좌의 적절한 평균 농도를 갖는 푸아송 확률 밀도

함수에 따라 생성된 각각의 유전자좌를 나타내는 난수에 의해 측정하였다.

[0209] 단계 2에서, 원하는 수의 웰, 예를 들어 384-웰 플레이트 실험에 대하여 384개 웰이 생성될 때까지 단계 1을 반복하였다. 염색체 1 단독 및 염색체 21 단독에 대해서만 양성인 웰의 수를 계수하였다. 정보제공 웰의 수는 염색체 중 하나에 대해 양성이되, 모두에 대해서는 아닌 웰의 총 수였다. 이후, 염색체 21에 대해 양성인 정보제공 웰의 비율 (P_r)을 계산하였다. 귀무가설 또는 대립가설을 수용하기 위한 관련 SPRT 곡선의 상한 및 하한을 SPRT 분석에 대해 앞선 섹션에서 기재된 것과 같이 계산하였다.

[0210] 단계 3에서, 정배수체 또는 삼염색체 21 태아를 갖는 임신 여성의 2개 시나리오 각각에 대해 5000회의 시뮬레이션을 수행하였다. 각각의 시뮬레이션은 임신 여성으로부터 얻어진 독립적인 생물학적 샘플로서 간주될 수 있었다. 표 1100에서, 정배수체 경우의 정확한 분류는 귀무가설이 수용되는 정배수체 경우를 의미하며, 정배수체 경우의 부정확한 분류는 대립가설이 수용되는 정배수체 경우를 의미한다. 유사하게, 대립가설이 수용되는 삼염색체 21 경우는 정확하게 분류된 것으로 간주되며, 귀무가설이 수용되는 삼염색체 21 경우는 부정확하게 분류된 것으로 간주되었다. 두 군 모두에 대하여, 미리-명시된 웰의 총 수를 시뮬레이션한 이후 귀무가설 또는 대립가설 모두가 수용되지 않는 경우에, 미분류된 것으로 간주되었다.

[0211] 단계 4에서, 단계 1 내지 3을 384개 내지 7680개 범위의 총 웰 수에서 10% 25%, 50% 및 100% 태아 DNA를 갖는 샘플에 대해 반복하였다.

[0213] II. 삼염색체 21 검출의 검증

[0214] **A. PLAC4에 대한 RNA-SNP**

[0215] 염색체 21 상의 PLAC4 유전자에 대한 rs8130833 SNP를 사용하여 디지탈 RNA-SNP의 실제적 실행가능성을 입증하였다 (문헌 [Lo, YMD et al. 2007 Nat Med 13, 218-223]). 2개 정배수체 및 2개 삼염색체 21 이형접합 태반으로부터의 태반 DNA 및 RNA 샘플을 분석하였다. 디지탈 RNA-SNP 프로토콜을 사용하되, 역전사 단계를 생략하였으며, 이에 따라 필수적으로 절차를 디지탈 DNA-SNP 분석으로 전환하여, 태반 DNA 샘플을 분석하였다. 정확한 경우 분류의 가능성과 정보제공 웰의 비율 사이에 균형을 맞추기 위해, 본 발명자들은 웰 당 임의의 유형의 하나의 대립유전자를 목적으로 하는 샘플을 희석하고, 96-웰 디지탈 PCR 분석에 의해 확인하였다. 이후, 384-웰 디지탈 RNA-SNP 실험을 하였다. P_r 및 m_r을 계산하고, 상기 m_r 값에 대한 SPRT 곡선을 데이다 해석을 위해 사용하였다.

[0216] 도 13a는 본 발명의 실시양태에 따른 정배수체 및 삼염색체 21 임신의 태반 조직에서의 디지탈 RNA-SNP 분석의 표 1300을 나타낸다. 질량 분광 검정에 의해 유전자형을 측정하였다. 실험적으로 얻어진 P_r이 분류불가능한 영역 아래에 있는 경우 "정배수체"로 지정하였고, 실험적으로 얻어진 P_r이 분류불가능한 영역 위에 있는 경우 삼염색체 21을 나타내는 "T21"로 지정하였다. T21, 삼염색체 21. 1회 384-웰 실험으로 DNA 및 RNA 샘플 모두를 사용하여 각각의 상기 경우를 정확하게 분류하였다.

[0217] 추가적으로, 본 발명자들은 정배수체 태아를 갖는 9명의 여성 및 삼염색체 21 태아를 갖는 4명의 여성으로부터의 혈장 RNA 샘플을 시험하였다. 도 13b는 본 발명의 실시양태에 따른 정배수체 및 삼염색체 21 임신 여성으로부터의 모체 혈장의 디지탈 RNA-SNP 분석의 표 1350을 나타낸다. 모든 경우를 정확하게 분류하였다. 한 삼염색체 21 경우 (M2272P)에 대한 초기 결과는 1회 384-웰 실험 이후 SPRT 곡선 사이의 분류불가능한 영역 내에 있었다. 따라서, 추가의 384-웰 실행을 수행하였다. 총 768개 웰로부터의 응집된 데이타로부터 신규 m_r 및 P_r 값을 계산하였으며, 상기 m_r 값을 기준으로 선택된 신규 세트의 SPRT 곡선을 사용하여 분류를 수행하였다. 이후, 상기 경우를 이수체로서 정확하게 스코어링하였다.

[0218] 본 발명자의 실험 및 시뮬레이션 데이타는 디지탈 RNA-SNP가 삼염색체 21 검출에 대한 효과적이며 정확한 방법임을 나타낸다. 모체 혈장 중 *PLAC4* mRNA가 태아로부터 순수하게 유래되기 때문에, 시험된 13개 모체 혈장 샘플 중 12개에 대하여, 오직 1회의 384-웰 디지탈 PCR 실험만이 정확한 분류를 위해 요구되었다. 이에 따라, 상기 균질한 실시간 디지탈 PCR-기반 방법은 RNA-SNP 분석을 위한 질량 분광법-기반 방법과는 또다른 방법을 제공한다. 태반-특이적 mRNA 전사체와는 별개로, 본 발명자들은 또한 모체 혈장에서의 다른 유형의 태아-특이적 핵산 종이 태아 염색체 이수성의 디지탈 PCR-기반 검출에 대해 사용될 수 있다는 것을 구상한다. 한 예는 최근에 후성 대립유전자 비 (EAR) 방법을 사용하여 삼염색체 18의 비-침습적 출생 전 검출에 사용된 태아 후성 마커이다 (문헌 [Tong YK et al. 2006 Clin Chem, 52, 2194-2202]). 따라서, 본 발명자들은 디지탈 EAR이 가능한 분

석 기술일 것이라고 예상한다.

[0220] **B. RCD**

[0221] 또한, 염색체 21 및 1에서의 파라로그 서열을 표적화하는 PCR 분석을 사용하여 삼염색체 21 검출에 대한 디지탈 RCD의 실제적 실행가능성을 연구하였다. 파라로그 유전자좌를 본원에서 예로서 사용하였다. 염색체 21 및 임의의 다른 참조 염색체에 대한 비-파라로그 서열 또한 RCD에 대해 사용할 수 있다. 2개 정배수체 및 2개 삼염색체 21 태반으로부터의 태반 DNA 샘플을 웰 당 염색체 중 하나로부터 대략 1개 표적 주형으로 희석시켰으며, 96-웰 디지탈 PCR 분석에 의해 확인하였다. 각각의 확인된 샘플을 384-웰 디지탈 RCD 실험에 의해 분석하였으며, P_r 및 m_r 값을 계산하였다. 디지탈 RCD에 대하여, 염색체 1 파라로그는 참조 주형이었다. 상기 m_r 값을 사용하여 데이타 해석을 위한 상응하는 세트의 SPRT 곡선을 선택하였다. 모든 태반 샘플을 도 14a에서 나타낸 것과 같이 정확히 분류하였다.

[0222] 모체 혈장에서의 태아 DNA의 시나리오와 같은 디지탈 RCD 방법을 사용하여 과량의 정배수체 DNA와 혼합된 삼염색체 21 DNA를 검출할 수 있는지를 입증하기 위해, 정배수체 모체 혈액 세포 DNA의 배경 중 50% 및 25%의 삼염색체 21 태반 DNA를 함유하는 혼합물을 분석하였다. 10개의 삼염색체 21 및 10개의 정배수체 경우로부터의 태반 DNA를 각각 동일한 양의 정배수체 모체 혈액 세포 DNA와 혼합하였으며, 이에 따라 20개의 50% DNA 혼합물을 제조하였다. 도 14b는 본 발명의 실시양태에 따른 50% 태아 DNA 혼합물의 RCD 분석에 대한 SPRT 해석을 예시하는 플롯 (1440)을 나타낸다. 유사하게는, 5개의 삼염색체 21 및 5개의 정배수체 경우로부터의 태반 DNA를 각각 3배 과량의 정배수체 모체 혈액 세포 DNA와 혼합하였으며, 이에 따라 10개의 25% DNA 혼합물을 제조하였다. 도 14c는 25% 태아 DNA 혼합물의 RCD 분석에 대한 SPRT 해석을 예시하는 플롯 1470을 나타낸다. 도 14b 및 14c에서 나타낸 것과 같이, 모든 정배수체 및 이수체 DNA 혼합물을 정확하게 분류하였다.

[0223] 각각의 샘플은 도 14b 및 14c에 표시된 것과 같이 여러 384-웰 디지탈 PCR 분석 이후 분류가능한 지점에 도달하였다. 50% DNA 혼합물에 대하여, 384-웰 플레이트의 수는 1개 내지 5개의 범위가 요구된다. 25% DNA 혼합물에 대하여, 384-웰 플레이트의 수는 1개 내지 7개의 범위가 요구된다. 매 384회의 디지탈 PCR 분석의 추가와 함께 정확하게 분류되는 경우의 누적 비율은 표 1200에서 제시된 컴퓨터 시뮬레이션에 의해 예상된 것과 같았다.

[0225] III. 디지탈 PCR을 이용한 방법

[0226] **A. 디지탈 RNA-SNP**

[0227] 모든 RNA 샘플을 먼저 써모스크립트(ThermoScript) 역전사효소 (인비트로젠(Invitrogen))를 사용하여 유전자-특이적 역전사 프라이머에 의해 역전사시켰다. 역전사 프라이머의 서열은 **5'-AGTATATAGAACCATGTTTAGGCCAGA-3'** (서열 1) (미국 아이오와주 코럴빌 소재의 인테그레이티드 DNA 테크놀러지스(Integrated DNA Technologies))였다. 디지탈 RNA-SNP를 위한 역전사된 RNA (즉, cDNA) 샘플 및 DNA 샘플 (예를 들어, 태반 DNA)의 후속 처리는 본질적으로 동일하였다. 디지탈 PCR 분석 전에, 먼저 DNA 및 cDNA 샘플을 프라이머 **5'-CCGCTAGGGTGTCTTTTAAGC-3'** (서열 2), **5'-GTGTTGCAATACAAAATGAGTTTCT-3'** (서열 3), 및 형광 프로브 **5'-(FAM)ATTGGAGCAAATTC(MGBNFQ)-3'** (서열 4) (미국 캘리포니아주 포스터 시티 소재의 어플라이드 바이오시스템즈) (여기서, FAM은 6-카르복시플루오레세인이고, MGBNFQ는 마이너 그루브 결합 비-형광 켄처임)로 이루어진, PLAC4에 대한 실시간 PCR 검정을 사용하여 정량화하였다.

[0228] 앰플리콘이 특정된 HPLC-정제된 단일 가닥 합성 DNA 올리고뉴클레오티드 (싱가포르 소재의 프롤리고(Proligo))의 계열 희석에 의해 보정 곡선을 제작하였다. 서열은 **5'-CGCCGCTAGGGTGTCTTTTAAGCTATTGGAGCAAATTCAAATTTGGCTTAAAGAA AAAGAAACTCATTTTGTATTGCAACACCAGGAGTATCCCAAGGGACTCG-3'** (서열 5)였다. 반응 부피 25 μL로 2X 태크만 유니버셜 PCR 마스타 믹스(TaqMan Universal PCR Master Mix) (어플라이드 바이오시스템즈)를 사용하여 반응을 설정하였다. 각각의 프라이머 400 nM 및 프로브 80 nM을 각각의 반응에 사용하였다. ABI PRISM 7900HT 서열 검출 시스템 (어플라이드 바이오시스템즈)에서, 반응을 50℃에서 2분 동안 개시하고, 이어서 95℃에서 10분 동안 진행하고, 95℃에서 15초 및 60℃에서 1분의 45주기를 수행하였다. 그 다음, 웰 당 대략 1개의 주형 분자에서 후속 디지탈 PCR 증폭이 수행될 수 있도록 DNA 또는 cDNA 샘플의 계열 희석에

착수하였다. 이러한 농도에서는, 반응 웰의 대략 37%가 음성 증폭을 나타낼 것으로 예측되며, 이는 96-웰 디지탈 실시간 PCR 분석을 수행함으로써 최초로 확인되었다. 이어서, 한 세트의 비-인트론 관통(spanning) 프라이머: 전방향 프라이머 5'-TTTGTATTGCAACACCATTTGG-3' (서열 6) 및 앞서 기재된 유전자-특이적 역전사 프라이머를 사용하여 384-웰 플레이트에서 디지탈 RNA-SNP 분석을 수행하였다.

[0229] PLAC4 서열 상 rs8130833 SNP의 2개의 대립유전자를 각각 표적으로 하는 2개의 대립유전자-특이적 태크만 프로브를 설계하였다. 그들의 서열은 G 및 A 대립유전자에 대하여 각각 5'- (FAM)TCGTCGTCTAACTTG(MGBNFQ)-3' (서열 7) 및 5'-(VIC)ATTCGTCATCTAACTTG(MGBNFQ)-3' (서열 8)였다. 반응 부피 5 μL로 2X 태크만 유니버셜 PCR 마스타 믹스를 사용하여 반응을 실정하였다. 각각의 반응은 1X 태크만 유니버셜 PCR 마스타 믹스, 각각의 프라이머 572 nM, 대립유전자-G-특이적 프로브 107 nM 및 대립유전자-A-특이적 프로브 357 nM을 함유하였다. 반응은 ABI PRISM 7900HT 서열 검출 시스템에서 수행하였다. 반응을 50℃에서 2분 동안 개시하고, 이어서 95℃에서 10분 동안 진행하고, 95℃에서 15초 및 57℃에서 1분의 45주기를 수행하였다. 반응 동안, SDS 2.2.2 소프트웨어 (어플라이드 바이오시스템즈)의 "절대 정량화" 응용 프로그램에 의해 형광 데이타를 수집하였다. 상기 소프트웨어는 기준선 및 임계값을 자동으로 계산하였다. A 또는 G 대립유전자에 대해 양성인 웰의 수를 기록하여 SPRT 분석에 제공하였다.

[0231] **B. 디지탈 RCD 분석**

[0232] 본 연구에 사용되는 모든 태반 및 모체 연막 DNA 샘플을 먼저 나노드롭 분광광도계 (미국 델라웨어주 윌밍턴 소재의 나노드롭 테크놀러지(NanoDrop Technology))에 의해 정량화하였다. 6.6 pg/세포의 전환을 사용하여 DNA 농도를 카피/μL로 전환하였다. 웰 당 대략 1개의 주형에 상응하는 DNA의 양을, DNA 샘플을 계열 희석함으로써 결정하고 96-웰 포맷에서 실시간 PCR 검정에 의해 확인하였는데, 여기서 웰의 대략 37%가 음성 증폭을 나타낼 것으로 예측되었다. 확인 플레이트에 대한 PCR 설정은 참조 염색체에 대한 프로브를 추가한 것만 제외하고는 이하에 기재된 것과 동일하였다. 디지탈 RCD 분석에서, 염색체 21 및 1 상의 파라로그 유전자좌 (문헌 [Deutsch, S. et al. 2004 J Med Genet 41, 908-915])를 먼저 전방향 프라이머 5'-GTTGTTCTGCAAAAAACCTTCGA-3' (서열 9) 및 역방향 프라이머 5'-CTTGGCCAGAAATACTTCATTACCATAT-3' (서열 10)에 의해 동시 증폭시켰다. 염색체 21 및 1 파라로그를 표적으로 하도록 2개의 염색체-특이적 태크만 프로브를 설계하였는데, 그들의 서열은 각각 5'-(FAM)TACCTCCAT AATGAGTAAA(MGBNFQ)-3' (서열 11) 및 5'-(VIC)CGTACCTCTGTAATGTGTAA(MGBNFQ)-3' (서열 12)였다. 각각의 반응은 1X 태크만 유니버셜 PCR 마스타 믹스 (어플라이드 바이오시스템즈), 각각의 프라이머 450 nM 및 각각의 프로브 125 nM을 함유하였다. 총 반응 부피는 5μL/웰이었다. 반응을 50℃에서 2분 동안 개시하고, 이어서 95℃에서 10분 동안 진행하고, 95℃에서 15초 및 60℃에서 1분의 50주기를 수행하였다. 모든 실시간 PCR 실험은 ABI PRISM 7900HT 서열 검출 시스템 (어플라이드 바이오시스템즈) 상에서 수행하였고, SDS 2.2.2 소프트웨어 (어플라이드 바이오시스템즈)의 "절대 정량화" 응용 프로그램에 의해 형광 데이타를 수집하였다. 디폴트 기준선 및 매뉴얼 임계값을 사용하였다. 염색체 21 또는 염색체 1에 대해 양성인 웰의 수를 기록하여 SPRT 분석에 제공하였다. SPRT에 의해 질환 분류가 가능해질 때까지 1개 이상의 384-웰 플레이트가 분석될 것이다.

[0234] <u>IV. 미세유체-기반 디지탈 PCR의 사용</u>

[0235] **A. 디지탈 RNA-SNP**

[0236] 본 실시예는 미세유체-기반 디지탈 PCR을 사용한 디지탈 PCR 분석의 성능을 증명한다. 플루이다임 바이오마크(Fluidigm BioMark)™ 시스템을 사용하는 이 접근법의 변형이 그에 제한되지 않는 예로서 본원에 예시되었다. 상기 시스템은 실행 1회당 9000회 초과의 디지탈 PCR을 수행할 수 있다.

[0237] 정배수체 또는 삼염색체 21 태아를 가진 임신 여성으로부터 태반 조직 및 모체 말초 혈액 샘플을 채취하였다. 프라이머 연장에 이은 질량 분광법에 의해 태반 DNA 샘플에서 PLAC4 유전자 상 rs8130833 SNP의 유전자형 결정을 수행하였다. 태반 및 모체 혈장 샘플로부터 RNA를 추출하였다.

[0238] 모든 RNA 샘플을 써모스크립트 역전사효소 (인비트로젠)를 사용하여 유전자-특이적 역전사 프라이머 (5'-AGTATATAGAACCATGTTTAGGCCAGA-3' (서열 13))에 의해 역전사시켰다. 태반 cDNA

샘플에 대해, 웰 당 대략 1개의 주형 분자에서 후속 디지탈 PCR 증폭이 수행될 수 있도록 계열 희석을 수행하였다.

[0239] 12.765 디지탈 어레이 (플루이다임)에 의해 바이오마크 시스템™ (플루이다임) 상에서 디지탈 PCR을 수행하였다. 각각의 디지탈 어레이는 12개의 샘플-검정 혼합물을 수용하기 위한 12개의 패널로 이루어져 있다. 각각의 패널을 7 nL 반응물/웰을 수행하기 위한 765개 웰로 추가 분할하였다. PLAC4 유전자 상의 rs8130833 SNP 영역을 전방향 프라이머 (5'-TTTGTATTGCAACACCATTTGG-3' (서열 14)) 및 앞서 기재된 유전자-특이적 역전사 프라이머에 의해 증폭시켰다. rs8130833 SNP의 2개의 대립유전자를 각각 표적으로 하는 2개의 대립유전자-특이적 태크만 프로브를 설계하였다. 그들의 서열은 G 및 A 대립유전자에 대하여 각각 5'-(FAM)TCGTCGTCTAACTTG(MGBNFQ)-3' (서열 15) 및 5'-(VIC)ATTCGTCATCTAACTTG(MGBNFQ)-3' (서열 16)였다. 반응 부피 10 μL로 2X 태크만 유니버셜 PCR 마스타 믹스를 사용하여 1개의 어레이 패널에 대한 반응을 설정하였다. 각각의 반응은 1X 태크만 유니버셜 PCR 마스타 믹스, 각각의 프라이머 572 nM, 대립유전자-G-특이적 프로브 53.5 nM, 대립유전자-A-특이적 프로브 178.5 nM 및 cDNA 샘플 3.5 μL를 함유하였다. 각각의 태반 cDNA 샘플에 대해서는 1개의 반응 패널이 사용된 반면, 각각의 모체 혈장 샘플에 대해서는 12개의 패널이 사용되었다. 샘플-검정 혼합물을 나노플렉스(NanoFlex)™ IFC 제어기 (플루이다임)에 의해 디지탈 어레이에 로딩하였다. 반응은 바이오마크™ 시스템에서 수행하였다. 반응을 50℃에서 2분 동안 개시하고, 이어서 95℃에서 10분 동안 진행하고, 95℃에서 15초 및 57℃에서 1분의 40주기를 수행하였다.

[0240] 1개의 정배수체 및 2개의 T21 이형접합체 태반으로부터의 태반 RNA 샘플을 765-웰 반응 패널에서 분석하였다. 각각의 샘플에 대해, A 또는 G 대립유전자 (둘 다는 아님)에 대해 양성인 것을 포함하는 정보제공 웰의 수를 계수하였다. 모든 정보제공 웰 중 과다표출된 대립유전자의 비율 (P_r)을 결정하였다. 이들 실행의 웰 당 정확한 평균 참조 주형 농도 (m_r)에 적당한 SPRT 곡선을 적용하여, 실험적으로 얻은 P_r이 정배수체를 지시하는지 T21 샘플을 지시하는지를 결정하였다. 도 15a에 나타낸 바와 같이, 상기 접근법을 사용하여 모든 RNA 샘플이 정확하게 분류되었다.

[0241] 정배수체 태아를 가진 4명의 여성 및 삼염색체 21 태아를 가진 1명의 여성으로부터 채취한 혈장 RNA 샘플을 추가로 시험하였다. 각각의 샘플을 12개의 765-웰 반응 패널, 즉 혈장 RNA 샘플 당 9180회의 반응으로 분석하였다. 도 15b는 상기 혈장 RNA 샘플의 12개 패널 각각에 대한 정보제공 웰의 수를 나타낸다. 표에 나타낸 바와 같이, 혈장 샘플 중 주형 농도는 인의의 1개의 반응 패널 중의 정보제공 웰의 수로는 SPRT 분류에 충분하지 않을 정도로 묽었다. 상기 샘플을 정배수체 샘플로 분류하기 전에, 3개의 반응 패널로부터의 정보제공 웰을 합해야 했다 (도 15c). 도 15c는 2 내지 12개의 패널로부터 종합한 데이타를 사용하여 모든 혈장 사례가 정확하게 분류될 수 있었음을 나타낸다.

[0242] 디지탈 PCR을 수행하기 위한 매뉴얼 방법과 비교해서, 상기 미세유체-기반 방법은 훨씬 더 빠르고 덜 노동 집약적이었다. 전체 과정이 2시간 반 내에 완료될 수 있었다.

[0244] **B. 삼염색체 18의 출생 전 검출을 위한 디지탈 RNA-SNP 분석**

[0245] 본 실시예에서는, 삼염색체 18 태아에 대한 다형성 대립유전자의 비율 불균형을 검출하기 위해, 염색체 18 상의 태반-발현 전사체인 세르핀 펩티다제 억제제 클레이드 B (오브알부민) 멤버 2 (SERPINB2) mRNA에 대한 디지탈 PCR-기반 대립유전자 식별 검정을 사용하였다. QIAamp DNA 미니 키트 (독일 힐덴 소재의 퀴아젠(Qiagen)) 및 TRIzol 시약 (미국 캘리포니아주 칼즈베드 소재의 인비트로젠)을 제조업자의 프로토콜에 기재된 바와 같이 각각 사용하여, 태반 조직 샘플로부터의 DNA 및 RNA 추출을 수행하였다. 추출한 태반 RNA 샘플을 DNase I (인비트로젠) 처리하여 오염된 게놈 DNA를 제거하였다. 앞서 기재된 바와 같은 매스어레이 컴팩트(MassARRAY Compact) (미국 샌디에고 소재의 시퀘놈(Sequenom))를 사용하여 호모지너스 매스익스텐드(Homogenous MassEXTEND) (hME) 검정에 의해 태반 조직 DNA 샘플에서 SERPINB2 유전자 상 rs6098 SNP의 유전자형 결정을 수행하였다.

[0246] 써모스크립트 역전사효소 (인비트로젠)를 사용하여 유전자-특이적 프라이머 5'-CGCAGACTTCTCACCAAACA-3' (서열 17) (미국 아이오와주 코릴빌 소재의 인테그레이티드 DNA 테크놀러지스)에 의해 태반 조직 RNA 샘플 상에서 SERPINB2 전사체에 대한 역전사를 수행하였다. 모든 cDNA 샘플을, 반응 웰 당 1개의 주형 분자의 평균 농도에서 후속 디지탈 PCR 증폭이 수행될 수 있도록 하는 농도로 희석하였다. 태크만 유니버셜 PCR 마스타 믹스 (미국 캘리포니아주 포스터 시티 소재의 어플라이드 바이오시스템즈) 및 바이오마크™ PCR 시약 (미국 샌프란시스코 소재의 플루이다임)을 사용하여 디지탈 PCR을

설정하였다. 전방향 프라이머 5'-CTCAGCTCTGCAATCAATGC-3' (서열 18) (인테그레이티드 DNA 테크놀러지스) 및 역방향 프라이머 (역전사에 사용된 유전자-특이적 프라이머와 동일한 것)는 600 nM의 농도로 사용하였다. SERPINB2 서열 상 rs6098 SNP의 A 또는 G 대립유전자를 표적으로 하는 2개의 테크만 프로브는 5'-(FAM)CCACAGGGAATTATTT(MGBNFQ)-3' (서열 19) 및 5'-(VIC)CCACAGGGGATTATTT(MGBNFQ)-3' (서열 20) (어플라이드 바이오시스템즈)였다. FAM은 6-카르복시플루오레세인이고, MGBNFQ는 마이너 그루브 결합 비-형광 켄처이며, 각각 300 nM 및 500 nM의 농도로 사용하였다. 각각의 샘플-시약 혼합물을 나노플렉스™ IFC 제어기 (플루이다임)를 사용하여 바이오마크™ 12.765 디지탈 어레이 상의 765개 반응 웰에 분할하였다. 분할 후, 어레이를 열 증폭 및 형광 검출을 위해 바이오마크™ 실시간 PCR 시스템 (플루이다임)에 넣었다. 반응을 50℃에서 2분 동안 개시하고, 95℃에서 5분 동안 지속한 후, 95℃에서 15초 및 59℃에서 1분의 45주기를 수행하였다. 증폭 후, 정보제공 웰 (A 또는 G 대립유전자 단독에 대해 양성인 것)의 수 및 대립유전자 둘 다에 대해 양성인 웰의 수를 계수하여 순차 확률비 검증 (SPRT) 분석에 제공하였다.

[0247] 이형접합 정배수체 태아에 대해, A 및 G 대립유전자는 태아 게놈에서 동등하게 표출되어야 하는 반면 (1:1), 삼염색체 18에 대해서는, 한 대립유전자의 추가 카피가 존재함으로써 태아 게놈 내에 2:1의 비율이 주어질 것이다. 상이한 샘플의 해석에 대한 일련의 SPRT 곡선을 생성시켰다. 이들 곡선은 분류에 필요한 정보제공 웰의 주어진 총 수 (x축)에 대한, 과다표출된 대립유전자에 대해 양성인 정보제공 웰의 예상 비율 P_r (y축)을 나타낸다. 각각의 샘플에 대해, 실험적으로 유도된 P_r을 예상 P_r 값과 비교하였다. 상부 곡선 위쪽의 샘플은 삼염색체 18로 분류되고, 하부 곡선 아래쪽의 샘플은 정배수체로 분류되었다. 두 곡선 사이의 공간은 분류불가능한 영역이다.

[0248] 태아 삼염색체 18의 검출을 위한 디지탈 RNA-SNP 분석의 실행가능성을 SERPINB2 유전자 상 rs6098 SNP를 사용하여 증명하였다. 먼저 이형접합체 사례를 확인하기 위해서, 질량 분광법에 의해 정배수체 및 삼염색체 18 태아를 갖는 대상체로부터의 태반 조직 DNA 샘플의 유전자형을 결정하였다. 9개의 정배수체 및 3개의 삼염색체 18 이형접합체 태반을 발견하여 디지탈 RNA-SNP 분석에 제공하였다. 각각의 샘플에 대해 P_r 및 m_r을 계산하고, 상기 m_r값에 대한 SPRT 곡선을 질환 분류를 위해 사용하였다. 도 16a에 나타낸 바와 같이, 모든 샘플이 정확하게 분류되었다. 삼염색체 18 태반에 대한 P_r값은 분류불가능한 영역의 위쪽인 반면, 정배수체 태반에 대한 P_r값은 상기 영역의 아래쪽에 속하였다.

[0249] m_r = 0.1, 0.2 및 0.3을 기준으로 한 SPRT 곡선을 갖는 샘플을 도 16b에 나타내었다. 이들 데이타는 디지탈 RNA-SNP 방법이 삼염색체 18 임신에 대한 매우 유용한 진단 도구임을 시사한다. 2개의 곡선은 분류불가능한 영역에 대한 경계선을 나타낸다. 상부 곡선 위쪽의 데이타 지점을 갖는 샘플은 이수체로 분류되고, 하부 곡선 아래쪽의 데이타 지점을 갖는 샘플은 정배수체로 분류되었다.

[0251] **C. 디지탈 RCD 분석**

[0252] 본 실시예는 미세유체-기반 디지탈 PCR을 사용한 디지탈 RCD 분석의 성능을 증명한다. 플루이다임 바이오마크™ 시스템을 사용하는 이 접근법의 변형이 그에 제한되지 않는 예로서 본원에 예시되었다. 상기 시스템은 실행 1회당 9000회 초과의 디지탈 PCR을 수행할 수 있다.

[0253] 정배수체 또는 삼염색체 21 (T21) 태아를 가진 임신 여성으로부터 태반 조직, 모체 혈액 세포 및 양수 샘플을 얻었다. 10개의 T21 및 10개의 정배수체 사례로부터의 태반 DNA를 각각 동량의 정배수체 모체 혈액 세포 DNA와 혼합함으로써 20개의 50% DNA 혼합물을 제조하였다. 혼합물 샘플 중 정확한 태아 비율을 확실히 하기 위해, 추출한 DNA를 먼저 260 nm에서 광학 밀도 (OD) 측정에 의해 정량화하였다. 그 다음, 12.765 디지탈 어레이 (플루이다임)를 사용하여 바이오마크™ 시스템 (플루이다임)에 의해 이들을 디지탈 정량화하였다. 샘플의 정량화 검정은 참조 염색체에 대한 프로브를 사용한 것만 제외하고는 이하에 기재된 것과 동일하였다.

[0254] 50% DNA 혼합물 및 양수 샘플 내 염색체 양을, 염색체 1 상에 위치한 유전자좌에 대한 비-다형성 염색체 21 유전자좌의 디지탈 PCR 분석에 의해 결정하였다. 먼저 염색체 21 및 1 상의 1쌍의 파라로그 유전자좌의 101-bp 앰플리콘을 전방향 프라이머 5'-GTTGTTCTGCAAAAAACCTTCGA-3' (서열 21) 및 역방향 프라이머 5'-CTTGGCCAGAAATACTTCATTACCATAT-3' (서열 22)에 의해 동시 증폭시켰다. 염색체 21 및 1 파라로그 사이를 구별하도록 2개의 염색체-특이적 테크만 프로브를 설계하였는데, 그들의 서열은 각각

5'-(FAM)TACCTCCATAATGAGTAAA(MGBNFQ)-3' (서열 23) 및 5'-(VIC)CGTACCTCTGTAATGTGTAA(MGBNFQ)-3' (서열 24)였다. 본원에서 파라로그 유전자좌는 단지 예로서 사용되었다. 환언하면, 비-파라로그 유전자좌도 이러한 분석에 대해 사용될 수 있다.

[0255] 삼염색체 18 (T18)을 검출하기 위한 디지탈 RCD 접근법의 용도를 증명하기 위해, 염색체 21 및 18 상의 파라로그 서열을 표적으로 하는 또다른 검정을 설계하였다. 먼저 염색체 21 및 18 상의 파라로그 유전자좌의 128-bp 앰플리콘을 전방향 프라이머 5'-GTACAGAAACCACAAACTGATCGG-3' (서열 25) 및 역방향 프라이머 5'-GTCCAGGCTGTGGGCCT-3' (서열 26)에 의해 동시 증폭시켰다. 염색체 21 및 18 파라로그 사이를 구별하도록 2개의 염색체-특이적 태크만 프로브를 설계하였는데, 그들의 서열은 각각 5'-(FAM)AAGAGGCGAGGCAA(MGBNFQ)-3' (서열 27) 및 5'-(VIC)AAGAGGACAGGCAAC(MGBNFQ)-3' (서열 28)였다. 본원에서 파라로그 유전자좌는 단지 예로서 사용되었다. 환언하면, 비-파라로그 유전자좌도 이러한 분석에 대해 사용될 수 있다.

[0256] 모든 실험은 12.765 디지탈 어레이 (플루이다임)를 사용하여 바이오마크™ 시스템 (플루이다임) 상에서 수행하였다. 반응 부피 10 μL로 2X 태크만 유니버셜 PCR 마스타 믹스 (어플라이드 바이오시스템즈)를 사용하여 1개의 패널에 대한 반응을 설정하였다. 각각의 반응은 1X 태크만 유니버셜 PCR 마스타 믹스, 각각의 프라이머 900 nM, 각각의 프로브 125 nM 및 50% 태반/모체 혈액 세포 DNA 샘플 3.5 μL를 함유하였다. 샘플/검정 혼합물을 나노플렉스™ IFC 제어기 (플루이다임)에 의해 디지탈 어레이에 로딩하였다. 검출을 위해 바이오마크™ 시스템 상에서 반응을 수행하였다. 반응을 50℃에서 2분 동안 개시하고, 이어서 95℃에서 10분 동안 진행하고, 95℃에서 15초 및 57℃에서 1분의 40주기를 수행하였다.

[0257] 정배수체 및 T21 50% 태반/모체 혈액 세포 DNA 샘플을 디지탈 어레이 상에서 chr21/chr1 검정에 의해 분석하였다. 각각의 샘플에 대해, chr21 또는 chr1 마커 단독 (둘 다는 아님)에 대해 양성인 웰을 포함하는 정보제공 웰의 수를 계수하였다. 모든 정보제공 웰 중 과다표출된 마커의 비율 (P_r)을 결정하였다. 디지탈 PCR 패널 중 어느 하나에 대한 웰 당 정확한 평균 참조 주형 농도 (m_r)에 적당한 SPRT 곡선을 적용하여, 실험적으로 얻은 P_r이 정배수체 샘플을 지시하는지 T21 샘플을 지시하는지를 결정하였다. 결정이 이루어질 수 있을 때까지, 분류되지 않은 채로 남아있는 샘플에 대한 여분의 패널로부터 데이타를 종합하였다. 도 17에 나타낸 바와 같이, 모든 50% 태반/모체 혈액 세포 DNA 샘플이 상기 접근법을 사용하여 필요한 1 내지 4개 패널 범위의 데이타에 의해 정확하게 분류되었다. 도 18에 나타낸 바와 같이, 정확한 분류를 위한 결정 경계선을 나타내기 위해 SPRT 곡선을 또한 플롯팅하였다.

[0258] 정배수체 태아를 가진 23명의 여성 및 T21 태아를 가진 6명의 여성으로부터 얻은 양수 샘플에 대한 RCD 분석을 추가로 적용하였다. 각각의 샘플을 단일 765-웰 반응 패널에서 chr21/chr1 검정에 의해 분석하였다. 도 19는 SPRT 분류 개요를 나타낸다. 도 19에 나타낸 바와 같이, 모든 29개 샘플이 정확하게 분류되었다. 따라서, 디지탈 RCD 방법은 출생 전 진단에 사용되는 다양한 종류의 샘플 유형, 예컨대 양수 및 융모막 융모 생검에서 추위성체 (문헌 [Levett LJ, et al. A large-scale evaluation of amnio-PCR for the rapid prenatal diagnosis of fetal trisomy. Ultrasound Obstet Gynecol 2001; 17: 115-8]) 또는 단일 뉴클레오티드 다형성 (SNP) (문헌 [Tsui NB, et al. Detection of trisomy 21 by quantitative mass spectrometric analysis of single-nucleotide polymorphisms. Clin Chem 2005; 51: 2358-62]) 마커 또는 실시간 비-디지탈 PCR (문헌 [Zimmermann B, et al. Novel real-time quantitative PCR test for trisomy 21. Clin Chem 2002; 48: 362-3])을 사용하는 삼염색체의 검출에 대한 대안적인 접근법이다.

[0259] T18 사례를 검출하기 위한 시도로, 3개의 정배수체 및 5개의 T18 태반 DNA 샘플 상에 chr21/chr18 검정을 적용하였다. 모든 정보제공 웰 중 과다표출된 마커의 비율 (P_r)을 계산하였다. 이들 모두는 1개의 T18 사례가 정배수체로 잘못 분류된 것을 제외하고는 정확하게 분류되었다. 결과를 도 20에 요약하였다.

[0261] <u>V. 질량 분광법 플랫폼에 대한 멀티플렉스 디지탈 RCD 검정의 이용</u>

[0262] 모체 혈장의 단위 부피 당 존재하는 혈장 DNA 분자의 수는 제한적이다 (문헌 [Lo YMD, et al. 1998 Am J Hum Genet 62, 768-7758]). 예를 들어, 임신 초기에 상염색체 유전자좌, β-글로빈 유전자의 중앙 모체 혈장 농도는 태아 및 모계 모두에 기인하여 986개 카피/mL인 것으로 밝혀진 바 있다 (문헌 [Lo YMD, et al. 1998 Am J

Hum Genet 62, 768-7758]). 7,680개의 분자를 포획하기 위해서, 대략 8 mL의 모체 혈장으로부터 추출된 DNA가 필요할 것이다. 약 15 mL의 모체 혈액으로부터 수득가능한 상기 부피의 혈장은 통상적인 실행의 한계이다. 그러나, 본 발명자들은 chr21 및 참조 염색체 표적의 다중 세트를 디지탈 RCD 분석을 위해 합할 수 있는 것으로 생각한다. 5쌍의 chr21 및 참조 염색체 표적에 대해, 단지 1.6 mL의 모체 혈장이 분석에 필요한 주형 분자의 수를 제공하기 위해 필요할 것이다. 멀티플렉스 단일 분자 PCR을 수행할 수 있었다. 이러한 멀티플렉스 단일 분자 분석의 견고성은 단일 분자 일배체형화에 대해 미리 입증된 바 있다 (문헌 [Ding, C. and Cantor, CR. 2003 Proc Natl Acad Sci U S A 100, 7449-7453]).

[0263] 일례에서, 태반 조직 및 모체 혈액 세포 샘플을 정배수체 또는 삼염색체 21 (T21) 태아를 가진 임신 여성으로부터 수득하였다. 5개의 정배수체 및 5개의 T21 태반 DNA 샘플을 각각 동일한 비율의 모체 혈액 세포 DNA와 혼합하여 50% 태아 DNA를 갖는 혈장 샘플을 모방한 10개의 DNA 혼합물을 생성하였다. 혼합물 샘플 중 정확한 태아 비율을 확실하게 하기 위해, 추출된 DNA를 먼저 260 nm에서의 광학 밀도 (OD) 측정으로 정량화하였다. 이어서, 이를 384-웰 포맷에서 실시간 PCR에 의해 디지탈 방식으로 정량화하였다. 샘플을 정량화하기 위한 검정은 상기 디지탈 RCD 분석의 예에 기재된 바와 동일하였다.

[0264] 50% 믹스 내 염색체 양을 염색체 1 상에 위치한 유전자좌에 대한 비-다형성 염색체 21 유전자좌의 디지탈 PCR 분석으로 결정하였다. 상기 방법은 디지탈 상대적 염색체 양 (RCD) 분석으로 칭해진다. 염색체 21 및 1 상의 한 쌍의 파라로그 유전자좌의 121-bp 앰플리콘 (각각의 프라이머 상에 10-mer 포함)을 전방향 프라이머 5'-ACGTTGGATGGTTGTTCTGCAAAAAACCTTCGA-3' (서열 29) 및 역방향 프라이머 5'-ACGTTGGATGCTTGGCCAGAAATACTTCATTACCATAT-3' (서열 30)를 사용하여 동시 증폭시켰다. 염색체 21 및 염색체 1 사이의 염기 차이를 표적으로 한 연장 프라이머를 고안하였고, 그의 서열은 5'-CTCATCCTCACTTCGTACCTC-3' (서열 31)이다.

[0265] T21 사례를 검출하는 멀티플렉싱 디지탈 PCR 검정의 유용성을 입증하기 위해서, 염색체 21 및 18 상의 파라로그 서열을 표적으로 한 또다른 디지탈 RCD 검정을 고안하였다. 염색체 21 및 18 상의 파라로그 유전자좌의 148-bp 앰플리콘 (각각의 프라이머 상에 10-mer 포함)을 전방향 프라이머 5'- ACGTTGGATGGTACAGAAACCACAAACTGATCGG-3' (서열 32) 및 역방향 프라이머 5'-ACGTTGGATGGTCC AGGCTGTGGGCCT-3' (서열 33)를 사용하여 동시 증폭시켰다. 염색체 21 및 염색체 18 사이의 염기 차이를 표적으로 한 연장 프라이머를 고안하였고, 그의 서열은 5'-ACAAAAGGGGGAAGAGG-3' (서열 34)이다.

[0266] 멀티플렉스 디지탈 RCD 분석은 프라이머 연장 프로토콜을 이용하여 수행되었다. PCR 반응은 진앰프(GeneAmp) PCR 코어 시약 키트 (어플라이드 바이오시스템즈)를 사용하여 반응 부피 5 μL로 설정하였다. 각각의 반응물은 1X 완충액 II, 2 mM $MgCl_2$, 200 μM dNTP 믹스, 0.2 U 앰플리태크 골드(AmpliTaq Gold), 각각 200 nM의 4종의 프라이머 및 50% DNA 믹스를 함유하였다. 상기 검정/샘플 혼합물을 384-웰 PCR 플레이트에 분배하고, 50℃에서 2분 동안 반응을 개시한 후, 95℃에서 10분 동안 진행하고, 95℃에서 15초 및 57℃에서 1분의 40회 주기를 수행하였다.

[0267] PCR 생성물을 쉬림프 알칼린 포스파타제 (SAP) 처리하여 혼입되지 않은 dNTP를 제거하였다. 상기 혼합물을 37℃에서 40분 동안, 이어서 85℃에서 5분 동안 인큐베이션하였다. 이어서, 프라이머 연장 반응을 수행하였다. 간단히 말해서, 연장 칵테일 중 chr21/chr1 검정으로부터의 771 nM의 연장 프라이머, chr21/chr18 검정으로부터의 1.54 μM의 연장 프라이머, 0.67 U 써모시퀴나제(Thermosequenase) (시쿼놈), 및 각각 64 μM의 ddCTP, ddGTP, dATP 및 dTTP를 SAP-처리된 PCR 생성물에 첨가하였다. 반응 조건은 94℃에서 2분, 이어서 94℃에서 5초, 50℃에서 5초 및 72℃에서 5초의 80회 주기였다. 16 μL의 물 및 3 mg의 클린 레진(Clean Resin) (시쿼놈)을 최종 처리(clean up)를 위해 연장 생성물에 첨가하였다. 상기 혼합물을 회전기에서 20 내지 30분 동안 혼합한 후, 361 g에서 5분 동안 원심 분리단계를 수행하였다. 15 내지 25 nL의 최종 생성물을 매스어레이 나노 분산기 S (시쿼놈)로 스펙트로칩(SpectroCHIP) (시쿼놈) 상에 분배하였다. 스펙트로칩으로부터의 데이타 획득을 매스어레이 분석기 컴팩트 질량 분광기 (시쿼놈)에서 수행하였다. 분석을 위해 질량 데이타를 매스어레이 타이퍼(MassARRAY Typer) (시쿼놈) 소프트웨어로 임포트하였다.

[0268] 5개의 정배수체 및 5개의 T21 50% 태반/모체 DNA 샘플을 이중 RCD 검정으로 분석하였다. 각 샘플에 대해, 오직 chr21 또는 chr1 또는 chr18 마커에만 양성인 웰을 포함하여 각각의 검정으로부터 정보제공 웰의 수를 계수하였다. 전체 정보제공 웰 중 chr21 마커의 비율 (P_r)을 각각의 RCD 검정에 대해 개별적으로 계산하였다. 이어서, 순차 확률비 검증 (SPRT)을 적용하여 P_r이 정배수체 샘플을 나타내는 지 또는 T21 샘플을 나타내는 지 결

정하였다. 이렇게 하여, 각각의 플레이트를 2회 계수함으로써 필요한 웰의 수를 감소시켰다.

[0269] chr21/chr1 검정을 통상 먼저 적용하였다. 샘플이 미분류 상태로 있는 경우, chr21/chr18 검정으로부터의 값을 추가 계산에 포함시킨 것이다. 여분의 플레이트를 결정이 이루어질 수 있을 때까지 미분류 상태로 있는 샘플에 사용하였다. 도 21에 나타낸 바와 같이, 모든 정배수체 50% 믹스 샘플을 단일 384-웰 플레이트를 사용하여 정확히 분류하였다. 어려 T21 사례는 정확한 분류를 위해 2개 이상의 플레이트를 필요로 하였다. 하나의 검정만을 이용한 경우, 분류 달성시 필요한 정보제공 웰의 수에 이르기 위해서 더 많은 플레이트가 필요할 것이다. 예를 들어, RCD 검정 중 어느 하나를 단독 이용한 경우에는 사례 N0230에 대한 데이타가 분류될 수 없었다. 그러나, 2개의 검정으로부터의 데이타를 합한 경우에는 정확한 분류가 이루어졌다. 이중 RCD 검정을 이용하지 않은 경우, 추가의 분석 플레이트가 필요할 것이다. 본 발명자들은 보다 높은 수준의 멀티플렉싱 검정으로 웰 수의 추가 감소를 예상할 것이다.

[0270] 또다른 예에서, 본 발명자들은 염색체 21, 및 다른 비-염색체-21 상염색체 상에 위치한 그의 상응하는 파라로그 파트너에 대한 4종의 상이한 앰플리콘을 표적으로 하는 4-플렉스 검정을 개발하였다. 이러한 4-플렉스 검정은 정배수체 및 삼염색체 21 임신 여성으로부터의 샘플의 디지탈 RCD 분석, 이어서 SPRT 분류에 사용되었다. 태반 샘플로부터의 DNA 추출은 퀴아앰프 조직 키트 (독일 힐덴 소재의 퀴아젠)를 사용하여 수행되었다.

[0271] 상기 연구에 사용된 모든 태반 및 모체 연막 DNA 샘플을 먼저 나노드롭 분광광도계 (미국 델라웨어주 윌밍턴 소재의 나노드롭 테크놀러지)로 정량화하였다. DNA 농도를 6.6 pg/세포의 전환을 이용하여 게놈 당량 (GE)/μL로 전환하였다. 웰 당 대략 1개의 주형에 해당하는 DNA의 양을 DNA 샘플의 계열 희석으로 결정하였다. 이러한 조건하에, 본 발명자들은 대략 37%의 웰이 음성 증폭을 나타낼 것으로 예상할 것이다. 멀티플렉스 디지탈 RCD 분석에서, 4 세트의 파라로그 서열 표적을 선택하였다: 염색체 21 및 1 상의 파라로그 유전자좌를 전방향 프라이머 5'-ACGTTGGATGTTGATGAAGTCTCATCTCTACTTCG-3' (서열 35) 및 역방향 프라이머 5'-ACGTTGGATGCAATAAGCTTGGCCAGAAATACT-3' (서열 36)를 사용하여 동시 증폭시켜 81 bp의 앰플리콘을 생성하였다. 염색체 21 및 7 상의 파라로그 유전자좌를 전방향 프라이머 5'-ACGTTGGATGGAATTTAAGCTAAATCAGCCTGAACTG-3' (서열 37) 및 역방향 프라이머 5'-ACGTTGGATGGTTTCTCATAGTTCATCGTAGGCTTAT-3' (서열 38)를 사용하여 동시 증폭시켜 82 bp의 앰플리콘을 생성하였다. 염색체 21 및 2 상의 파라로그 유전자좌를 전방향 프라이머 5'-ACGTTGGATGTCAGGCAGGGTTCTATGCAG-3' (서열 39) 및 역방향 프라이머 5'-ACGTTGGATGAGGCGGCTTCCTGGCTCTTG-3' (서열 40)를 사용하여 동시 증폭시켜 101 bp의 앰플리콘을 생성하였다. 염색체 21 및 6 상의 파라로그 유전자좌를 전방향 프라이머 5'-ACGTTGGATGGCTCGTCTCAGGCTCGTAGTT-3' (서열 41) 및 역방향 프라이머 5'-ACGTTGGATGTTTCTTCGAGCCCTTCTTGG-3' (서열 42)를 사용하여 동시 증폭시켜 102 bp의 앰플리콘을 생성하였다. 각각의 반응물은 10X 완충액 II (어플라이드 바이오시스템즈), $MgCl_2$ 및 각각 100 nM의 프라이머를 함유하였다.

총 반응 부피는 5 μL/웰이었다. 95℃에서 5분 동안 반응을 개시한 후, 95℃에서 30초, 62℃에서 30초 및 72℃에서 30초의 45회 주기를 수행하고, 72℃에서 7분 동안 최종 연장을 수행하였다. 모든 통상적인 PCR 증폭은 진앰프 PCR 시스템 9700 (어플라이드 바이오시스템즈) 상에서 수행되었다. 혼입되지 않은 뉴클레오티드를 쉬림프 알칼린 포스파타제 (SAP) 처리로 비활성화시켰다. 각각의 반응물은 10X SAP 완충액 (시쿼놈) 및 SAP 효소 (시쿼놈)를 함유하였다. 2 μL의 SAP 믹스를 각각의 PCR에 첨가하였다. SAP 반응물을 37℃에서 40분 동안 및 85℃에서 5분 동안 인큐베이션하였다. SAP 처리 후, iPLEX 골드(Gold) 키트 (시쿼놈)를 사용하여 PCR 생성물 상에서 프라이머 연상 반응을 수행하였다. 염색체 21 및 1 상의 파라로그 유전자좌에 대한 파라로그 서열 미스매치 (PSM)는 연장 프라이머 5'-GTCTCATCTCTACTTCGTACCTC-3' (서열 43)에 의해 검사되었다. 염색체 21 및 7 상의 파파로그 유전자좌에 대한 PSM은 연장 프라이머 5'-TTTTACGCTGTCCCCATTT-3' (서열 44)에 의해 검사되었다. 염색체 21 및 2 상의 파라로그 유전자좌에 대한 PSM은 연장 프라이머 5'-GGTCTATGCAGGAGCCGAC-3' (서열 45)에 의해 검사되었다. 염색체 21 및 6 상의 파라로그 유전자좌에 대한 PSM은 연장 프라이머 5'-TGGGCGCGGGAGCGCACTTCGCTGG-3' (서열 46)에 의해 검사되었다. 각각의 반응물은 10X iPLEX 완충액 (시쿼놈), iPLEX 종결 믹스 (시쿼놈), iPLEX 효소 (시쿼놈), 및 1.03 μM로 사용된 염색체 21 및 6에 대한 PSM용 연장 프라이머를 제외하고는 각각 343 nM의 연장 프라이머를 함유하였다. 2 μL의 iPLEX 믹스를 5 μL의 PCR 생성물에 첨가하였다. 200-쇼트-사이클 프로그램에 따라 iPLEX 반응을 순환시켰다. 간단히 말해서, 샘플을 먼저 94℃에서 35초 동안 변성시킨 후, 52℃에서 5초 동안 어닐링하고, 80℃에서 5초 동안 연장하였다. 어닐링 및 연장 주기를 총 5회의 주기 동안 4회 이상 반복하고, 이어서 94℃, 5초의 변성 단계로 역 순환시킨 후, 다시 5회 주기의 어닐링 및 연장 루프를 수행하였다. 1회의 변성 단계와 함께 5회의 어닐링 및 연장 주기를 39회 반복하여 총 40회에 이르게 하였다. 72℃에서 3분 동안 최종 연장을 수행하였다. iPLEX 반응 생성물을 16 μL의 물로 희석하고, 각각의 PCR에 대해 6 mg의 수지로 탈염시켰다. 384-웰 플레이트를 1600 g에서 3분 동안 원심분리한

후, 스펙트로칩 (시쿼놈) 및 매트릭스-보조 레이저 탈착/이온화 비행시간형 (MALDI-TOF) 질량 분광법 MS 분석기 (시쿼놈) 상에 분배하였다.

[0272] 염색체 21에 대해서만 또는 4개의 검정 각각에 대한 참조 염색체에 대해서만 양성인 웰의 수를 독립적으로 기록하였다. 각각의 검정에 있어서, 염색체 21 및 참조 염색체에 대한 푸아송 교정된 분자 수를 계산하였다. 염색체 21에 대한 푸아송 교정된 분자 수의 합 뿐만 아니라 전체 4개의 검정으로부터의 푸아송 교정된 참조 염색체 수의 합을 계산하였고, 이는 4-플렉스 검정에 대한 정보제공 계수로 간주되었다. P_r 값은 4-플렉스 검정에 대한 염색체 21 계수를 4-플렉스 검정에 대한 염색체 21 및 참조 염색체 계수의 합으로 나눈 값이었다. 실험적으로 유도된 P_r 값을 SPRT 분석에 적용하였다. SPRT에 의해 질환 분류가 가능해질 때까지 1개 이상의 384-웰 플레이트를 분석할 것이다. 총 2개의 50% 정배수체 태반 게놈 DNA/50% 모체 연막 DNA 믹스 및 2개의 50% 삼염색체 21 태반 게놈 DNA/50% 모체 연막 DNA 믹스를 분석하였다.

[0273] 실험적으로 유도된 P_r 값을 귀무가설 또는 대립가설 시험시의 예상 P_r 값과 비교할 것이다. 다르게는, 주어진 수의 정보제공 계수에 대한 P_r이 질환 분류에 필요한 통계적 신뢰 수준에 도달하지 못한 경우, 귀무가설 또는 대립가설은 모두 수용될 수 없을 것이다. 이러한 경우는 보다 많은 데이타가 이용가능해질 때까지 분류불가능인 것으로 간주되었다.

[0274] 각 샘플의 결과 및 SPRT 분류를 도 22a 및 22b에 표로 작성하였다. 2개의 정배수체 샘플은 SPRT 분류에 도달할 수 있기 전에 2개 및 5개의 384-웰 멀티플렉스 디지탈 RCD 분석을 필요로 하였다. 4-플렉스 검정의 개별 구성으로부터의 데이타는 어떤 것도 SPRT에 의한 질환 분류가 가능하지 않았다. 삼염색체 21 샘플은 둘 다 단지 하나의 384-웰 멀티플렉스 디지탈 RCD 분석으로 각각 정확히 분류되었다. 유사하게, 4-플렉스 검정의 개별 구성으로부터의 데이타는 어떤 것도 SPRT에 의한 질환 분류가 가능하지 않았다. 그러나, 4-플렉스 검정으로부터의 혼합 계수는 정확한 SPRT 분류를 가능하게 하였다. 이들 데이타는, 멀티플렉스 디지탈 RCD의 이용에 의해 정보제공 계수의 유효 수가 단일-플렉스 디지탈 RCD 검정의 이용과 비교하여 주어진 수의 수행된 디지탈 PCR 분석에 대해 상당히 증가하였음을 예증하였다.

[0276] VI. 디지탈 후성 상대적 염색체 양의 사용

[0277] 여기서, 본 발명자들은, 염색체 이수성에 관여하는 염색체 (즉, 삼염색체 21에서 염색체 21) 및 참조 염색체 상의 태아-특이적 DNA 메틸화 패턴 또는 다른 후성 변화를 나타내는 후성 마커를 디지탈 PCR 분석에 적용하는, 소위 디지탈 후성 상대적 염색체 양 (디지탈 ERCD)이라 불리는 접근법을 약술한다. 정상 태아를 가진 임신 여성으로부터 추출된 혈장 DNA 내 참조 염색체 후성 마커에 대해 양성인 웰 수에 대한 염색체 21 후성 마커에 대해 양성인 웰 수의 비율은 참조 범위를 제공할 것이다. 상기 비율은 태아가 삼염색체 21을 갖는 경우 증가할 것으로 예상될 것이다. 하나 초과의 염색체 21 마커 및 하나 초과의 참조 염색체 마커를 본 분석에서 사용할 수 있다는 것이 당업자에게 명백하다.

[0278] 태아 (태반)-특이적 메틸화 패턴을 나타내는 염색체 21 상의 유전자의 일례는 완전카르복실라제 합성효소 (Holocarboxylase synthetase, HLCS) 유전자이다. HLCS는 태반에서 과-메틸화되지만, 모체 혈액 세포에서는 저-메틸화되며, 이는 본원에 참조로 포함된 미국 특허 출원 제11/784499호에 포함된다. 태아 (태반)-특이적 메틸화 패턴을 나타내는 참조 염색체 상의 유전자의 일례는 염색체 3 [10] 상의 RASSF1A 유전자이다. RASSF1A는 태반에서 과-메틸화되지만, 모체 혈액 세포에서는 저-메틸화되며, 본원에 참조로 포함된 미국 특허 출원 제11/784501호를 참조한다.

[0279] 모체 혈장을 이용한 태아 내 삼염색체 21의 디지탈 PCR 검출에 대한 과-메틸화된 HLCS 및 과-메틸화된 RASSF1A의 적용에서, 모체 말초 혈액을 먼저 수집한다. 이어서, 상기 혈액을 원심분리하고, 혈장을 수집한다. 이어서, 혈장으로부터의 DNA를 당업계에 널리 공지된 기술, 예컨대 퀴아앰프 블러드(Blood) 키트 (퀴아젠)를 사용하여 추출한다. 이어서, 혈장 DNA를 1종 이상의 메틸화-감수성 제한 효소, 예컨대 HpaII 및 BstUI를 사용하여 소화시킨다. 이들 메틸화-감수성 제한 효소는 태아의 과-메틸화된 서열을 무손상 상태로 두는 반면, 이들 유전자의 모체, 비-메틸화된 버젼은 절단할 것이다. 이어서, 소화된 혈장 DNA 샘플을, 제한 효소 처리되었지만 무손상인 HLCS 또는 RASSF1A 서열의 분자가 반응 웰 당 평균 대략 0.2 내지 1개 검출될 정도로 희석한다. 2개의 실시간 PCR 시스템을 사용하여 희석된 DNA를 증폭시킬 것이며; 하나는 2개의 프라이머, 및 HLCS 유전자에 특이적인 1개의 태크만 프로브 (서열이 비-메틸화된 경우 제한 효소(들)에 의해 절단될 영역 포함)를 사용하고; RASSF1A 유전자에 관한 다른 하나는 유사하게 2개의 프라이머 및 1개의 태크만 프로브를 사용한다. 후자의 RASSF1A 프라이머/프로브 세트와 관련하여, 일례가 문헌 [Chan et al 2006, Clin Chem 52, 2211-2218]에 기재

되어 있다. HLCS 및 RASSF1A 표적에 관한 태크만 프로브는 상이한 형광 리포터, 예컨대 각각 FAM 및 VIC를 가질 것이다. 이어서, 384-웰 플레이트를 사용하여 디지탈 PCR 실험을 수행한다. HLCS에 대해서만 양성으로 스코어링된 웰의 수 및 RASSF1A에 대해서만 양성으로 스코어링된 웰의 수를 계수할 것이며, 이들 계수의 비율을 구할 것이다. HLCS:RASSF1A 비율은 정상 정배수체 태아를 가진 임신 여성과 비교하여 삼염색체 21 태아를 가진 임신 여성으로부터 얻은 모체 혈장에 대해 보다 높을 것으로 예상될 것이다. 과다표출의 정도는 디지탈 PCR 수행시 웰 당 평균 참조 주형 농도에 따라 달라질 것이다.

[0280] 이들 결과를 스코어링 하는 다른 방법, 예를 들어 RASSF1A에 대해 동시에 양성인 것에 관계 없이 HLCS에 대해 양성인 웰의 수를 계수하고, 반대로 HLCS에 대해 동시에 양성인 것에 관계 없이 RASSF1A에 대해 양성인 웰의 수를 계수하는 방법이 가능할 것이다. 또한, 비율을 계산하는 대신 HLCS 및 RASSF1A 계수의 총 수 또는 이들의 차이를 이용하여 태아의 삼염색체 21 상태를 나타낼 수 있었다.

[0281] 플레이트에서 디지탈 PCR을 수행하는 것 외에도, 디지탈 PCR의 다양한 변형, 예를 들어 미세유체 칩, 나노리터 PCR 마이크로플레이트 시스템, 에멀젼 PCR, 폴로니 PCR, 롤링-써클 증폭법, 프라이머 연장법 및 질량 분광법 등을 이용할 수 있다는 것이 당업자에게 명백할 것이다. 디지탈 PCR의 이러한 변형은 단지 예로서 제시되며, 제한적이지 않다.

[0282] 실시간 PCR 외에도, 질량 분광법과 같은 방법을 이용하여 디지탈 PCR 결과를 스코어링할 수 있다는 것이 당업자에게 명백할 것이다.

[0283] HLCS 및 RASSF1A의 태아 및 모체 버전을 식별하기 위해 메틸화-감수성 제한 효소를 사용하는 것 외에도, 메틸화 상태를 확인하기 위한 다른 방법, 예를 들어 바이설파이트 변형법, 메틸화-특이적 PCR, 메틸화된 시토신에 대한 항체를 사용하는 면역침강법, 질량 분광법 등을 적용할 수도 있음이 당업자에게 명백할 것이다.

[0284] 또한, 본 특허 출원의 상기 예 및 기타 예에서 설명된 접근법을 모체 뇨, 양수, 자궁경관 세척액, 융모막 융모, 모체 타액 등을 비롯한 태아 DNA가 발견될 수 있는 기타 체액에서도 사용할 수 있음이 당업자에게 명백할 것이다.

[0286] VII. 에멀젼 PCR 및 기타 전략을 이용한 대규모 병렬 게놈 서열분석

[0287] 여기서, 본 발명자들은, 핵산 분자의 디지탈 판독이 모체 혈장 중 태아 염색체의 이수성, 예를 들어 삼염색체 21을 검출하기 위해 사용될 수 있기 때문에 핵산 분자의 디지탈 판독의 또다른 예를 기재할 것이다. 태아 염색체 이수성은 염색체 또는 염색체 영역의 비정상적 용량(들)으로부터 기인한다. 비침습성 시험법이 감수성 및 특이성이 높기 때문에 오진을 최소화하기에 바람직하다. 그러나, 태아 DNA는 낮은 절대 농도로 존재하고, 모체 혈상 및 혈청 중의 모든 DNA 서열에 대해 소량의 부분으로 나타난다. 따라서, 디지탈 PCR 샘플링 표적화에 특이적인 유전자좌의 갯수는 동일한 생물학적 표본 내에서 무한히 증가될 수 없다. 따라서, 많은 세트의 특이적 표적 유전자좌를 분석함으로써, 디지탈 PCR 샘플링 수행 횟수를 증가시키지 않으면서도 표본으로부터 얻을 수 있는 데이타의 양을 증가시킬 수 있다.

[0288] 따라서, 본 실시양태는 모체 배경 핵산을 함유한 생물학적 샘플 중에 소량의 집단으로서 존재하는 제한된 양의 태아 핵산으로부터 추론될 수 있는 유전 정보의 양을 최대화함으로써 태아 염색체의 이수성을 비-침습적으로 검출한다. 한 측면에서, 얻어진 유전 정보의 양은 정확한 진단을 위해서는 충분하지만, 요구되는 생물학적 샘플의 투입량 및 비용을 포괄할 정도로 지나치게 과량인 것은 아니다.

[0289] 454 플랫폼 (로슈) (문헌 [Margulies, M. et al. 2005 Nature 437, 376-380]), 일루미나 게놈 분석기 (또는 솔렉사 플랫폼) 또는 솔리드 시스템 (어플라이드 바이오시스템즈) 또는 헬리코스 트루 싱글 몰레큘 DNA 서열분석 기술 (문헌 [Harris TD et al. 2008 Science, 320, 106-109]), 퍼시픽 바이오사이언시즈의 단일 분자, 실시간 (SMRT™) 기술, 및 나노기공 서열분석 (문헌 [Soni GV and Meller A. 2007 Clin Chem 53: 1996-2001]) 상에서 달성가능한 것과 같은 대규모 병렬 서열분석은 고도로 다중화된 표본으로부터 단리된 많은 핵산 분자를 병렬 방식으로 서열분석한다 (문헌 [Dear Brief Funct Genomic Proteomic 2003; 1: 397-416]). 이들 각각의 플랫폼은, 클로닝에 의해 확장된 또는 비-증폭된 핵산 질편의 단일 분자를 서열분석한다.

[0290] 대다수의 서열분석 판독이, 수십만 내지 수백만 또는 더 가능하게는 수백만 또는 수백억 정도로, 각각의 진행에서 각각의 샘플로부터 생성되기 때문에, 얻어진 서열분석된 판독은 원래 표본 중 핵산 종 혼합물의 대표적인 프로파일을 형성한다. 예를 들어, 서열분석된 판독의 일배체형, 전사체 및 메틸화 프로파일은 원래 표본의 그것과 유사하다 (문헌 [Brenner et al Nat Biotech 2000; 18: 630-634]; [Taylor et al Cancer Res 2007; 67: 8511-8518]). 각각의 표본으로부터 서열을 대량 샘플링하기 때문에, 핵산 풀(pool)의 서열분석으로부터 몇배의

적용범위 또는 고 중복성으로 생성된 것과 같은 동일한 서열의 수는 또한 원래 샘플 중 특정 핵산 종의 수 또는 유전자위 수에 대한 좋은 정량적 표출이다.

[0291] 한 실시양태에서, 랜덤 서열분석은 임신 여성의 혈장 중에 존재하는 DNA 절편에 대해 수행되고, 본래 태아 또는 모계로부터 유래되는 게놈 서열을 얻는다. 랜덤 서열분석은 생물학적 샘플에 존재하는 핵산 분자의 랜덤 부분을 샘플링 (서열분석)하는 것을 포함한다. 서열분석이 랜덤하게 이루어지기 때문에, 핵산 분자 (및 따라서 게놈)의 상이한 하위세트 (분획)를 각각의 분석에서 서열분석할 수 있다. 본 실시양태는 이러한 하위세트가 샘플마다 및 심지어 동일한 샘플을 사용하여 이루어질 수 있는 분석마다 다양한 경우라 해도 작용할 것이다. 분획의 예는 게놈의 약 0.1%, 0.5% 또는 1%이다. 다른 실시양태에서, 분획은 이들 값 중 적어도 하나이다.

[0292] 이어서, 바이오인포메틱스 절차를 이용하여 이들 DNA 서열 각각을 인간 게놈에 위치시킬 수 있다. 상기 서열의 비율은 이들이 인간 게놈의 반복 영역 또는 개인간 변이, 예를 들어 카피수 변이(copy number variation)에 적용되는 영역에 존재하기 때문에 추후 분석으로부터 감산될 수 있다. 이에 따라, 관심 염색체 및 하나 이상의 다른 염색체의 양을 측정할 수 있다.

[0293] 한 실시양태에서, 염색체 이수성, 예를 들어 염색체 21 또는 염색체 18 또는 염색체 13에 잠재적으로 포함되는 염색체의 파라미터 (예를 들어, 분획 표출)는 예를 들어 바이오인포메틱스 절차의 결과로부터 계산될 수 있다. 분획 표출은 모든 서열 (예를 들어, 모든 염색체의 일부 측정치) 또는 특정 하위세트의 염색체 (예를 들어, 시험된 것과는 다른 염색체)의 양을 기준으로 얻어질 수 있다.

[0294] 이후, 상기 분획 표출은 정상 태아 (즉, 정배수체)를 포함하는 임신에서 확립된 참조 범위와 비교한다. 일부 변형된 절차에서, 참조 범위를 특정 모계 혈장 샘플 중 태아 DNA의 분획 농도 (f)에 따라 조정하는 것이 가능하다. f 값은 태아가 남성인 경우 Y 염색체에 대해 맵핑가능한 서열을 이용하여 서열분석 데이타세트로부터 측정될 수 있다. f 값은 또한 별도 분석에서, 예를 들어 태아 후성 마커를 사용하여 (문헌 [Chan KCA et al 2006 Clin Chem 52, 2211-8]) 또는 단일 뉴클레오티드 다형성 분석으로부터 측정될 수 있다.

[0295] 한 측면에서, 표본 중의 핵산 풀을 100% 미만의 게놈 적용범위에서 서열분석하고, 포획된 핵산 분자의 비율 중에서, 대부분의 각각의 핵산 종만을 일단 서열분석하는 경우에, 특정 유전자위 또는 염색체의 용량 불균형이 또한 정량적으로 측정될 수 있다. 다시 말하면, 유전자위 또는 염색체의 용량 불균형은 표본의 모든 맵핑가능한 서열분석된 태그 중 상기 유전자위의 표출 백분율로부터 추론된다.

[0296] 대규모 병렬 게놈 서열분석 접근법에 대한 한 측면에서, 모든 염색체로부터의 대표적인 데이타는 동시에 생성될 수 있다. 특정 절편의 기원은 해당 시간 전에 선택되지 않는다. 서열분석은 랜덤하게 이루어지고, 이어서 데이타베이스 조사를 수행하여 상기 특정 절편이 어디로부터 유래되었는지 알 수 있다. 이것은 염색체 21로부터의 특정 절편 및 염색체 1로부터의 또다른 절편을 증폭시키는 상황과 대비된다.

[0297] 한 예에서, 이러한 서열의 부분은 상기 설명 예에서 염색체 21과 같은 이수성에 포함되는 염색체로부터 유래할 것이다. 이러한 서열분석 실행으로부터 다른 서열이 다른 염색체로부터 유래될 것이다. 다른 염색체와 비교되는 염색체 21의 상대적 크기를 고려함으로써, 참조 범위 내에서, 상기 서열분석 실행으로부터의 염색체 21-특이적 서열의 정규화 도수를 얻을 수 있다. 태아가 삼염색체 21을 갖는 경우, 서열분석 실행으로부터의 염색체 21-유래된 서열의 정규화 도수는 증가할 것이고, 따라서 삼염색체 21이 검출된다. 정규화 도수의 변화 정도는 분석된 샘플 중 태아 핵산의 분획 농도에 따라 좌우될 것이다.

[0298] 한 실시양태에서, 본 발명자들은 인간 게놈 DNA 및 인간 혈장 DNA 샘플의 단일-말단 서열분석을 위해 일루미나 게놈 분석기를 사용하였다. 일루미나 게놈 분석기는 유동 셀로 명명된 고체 표면 상에서 포획된, 클로닝에 의해 확장된 단일 DNA 분자를 서열분석한다. 각각의 유동 셀은 8개의 개별 표본 또는 표본 풀의 서열분석을 위한 8개 레인을 포함한다. 각각의 레인은 오직 인간 게놈 서열의 30억개의 염기쌍의 분획인 약 200Mb의 서열을 생성할 수 있다. 각각의 게놈 DNA 또는 혈장 DNA 샘플은 유동 셀의 한 레인을 이용하여 서열분석하였다. 생성된 짧은 서열 태그를 인간 참조 게놈에 대해 정렬시키고, 염색체 기원을 기록하였다. 각각의 염색체에 대해 정렬된 서열분석된 개별 태그의 총 갯수를 인간 참조 게놈 또는 비질환 대표 표본으로부터 예상되는 바와 같은 각각의 염색체의 상대적인 크기와 비교하여 표로 만들었다. 이렇게 하여, 염색체 획득 또는 소실을 확인하였다.

[0299] 기재된 접근법은 현재 기재된 유전자/염색체 용량 전략의 하나의 예시일 뿐이다. 별법으로, 페어링 말단의 서열분석을 수행할 수 있다. 문헌 [Campbell et al, Nat Genet 2008; 40: 722-729]에 기재된 바와 같이 참조 게놈에서 예상되는 것으로부터 서열분석된 절편의 길이를 비교하는 대신, 정렬된 서열분석된 태그의 갯수를 카운팅하고, 염색체 위치에 따라 분류하였다. 염색체 영역 또는 전체 염색체의 획득 또는 소실은 태그 카운트를 참

조 게놈에서의 예상되는 염색체 크기 또는 비질환 대표 표본의 예상되는 염색체 크기와 비교함으로써 측정하였다.

[0300] 또다른 실시양태에서, 실행 시 서열분석되는 핵산 풀의 분획은 서열분석 전에 추가로 하위 선택된다. 예를 들어, 올리고뉴클레오티드 어레이와 같이 혼성화를 기초로 하는 기술을 이용하여, 특정 염색체, 예를 들어 잠재적 이수체 염색체 및 시험된 이수성에 포함되지 않는 다른 염색체(들)로부터의 핵산 서열을 먼저 하위 선택할 수 있다. 또다른 예는 샘플 플로부터의 핵산 서열의 특정 하위-집단이 서열분석 이전에 하위 선택되거나 또는 풍부화된다는 것이다. 예를 들어, 모체 혈장 중 태아 DNA 분자가 모체 배경 DNA 분자보다 더 짧은 절편으로 구성되어 있는 것으로 보고되어 있다 (문헌 [Chan et al Clin Chem 2004; 50: 88-92]). 따라서, 분자 크기에 따라 샘플 내 핵산 서열을 분획화하기 위해 당업자들에게 공지된 하나 이상의 방법, 예를 들어 겔 전기영동 또는 크기 배제 컬럼, 또는 미세유체-기반 접근법에 의한 방법을 사용할 수 있다. 그러나, 별법으로, 모체 혈장 중의 무-세포 태아 DNA 분석 예에서는, 태아 핵산 부분이 모체 배경을 억제하는 방법에 의해, 예컨대 포름알데히드의 첨가에 의해 풍부화될 수 있나 (문헌 [Dhallan et al JAMA 2004; 291 : 1114-9]).

[0301] 로슈 454 플랫폼, 어플라이드 바이오시스템즈 솔리드 플랫폼, 헬리코스 트루 싱글 몰레큘 DNA 서열분석 기술, 퍼시픽 바이오사이언시즈의 단일 분자, 실시간 (SMRT™) 기술 및 나노기공 서열분석에 의해서와 같은 다른 단일 분자 서열분석 전략이 본 출원에서 유사하게 사용될 수 있나.

[0302] 결과 및 (예를 들어, 서열분석 및 파라미터 계산을 위한) 추가 논의의 예는 공동 출원된 출원 "게놈 서열분석을 사용한 태아 염색체 이수성 진단" (대리인 관리 번호 016285-005220US)에서 발견될 수 있고, 이것은 본원에 참고로 포함된다. 분별값을 측정하기 위해 본원에 기재된 방법은, 예를 들어 본 섹션에 기재된 바와 같이, 반응물을 서열분석하는 경우 실행될 수 있음을 주목한다.

[0303] 모체 혈장 중 태아 DNA의 분획 농도의 측정은 또한 서열분석 진행과 별도로 수행될 수 있다. 예를 들어, Y 염색체 DNA 농도는 실시간 PCR, 미세유체 PCR 또는 질량 분석기를 사용하여 미리 측정할 수 있다. 사실, 태아 DNA 농도는 여성 태아에 적용할 수 있는 Y 염색체 이외의 유전자좌를 사용하여 측정될 수 있다. 예를 들어, 챈 등(Chan et al)은 태아-유래된 메틸화된 RASSF1A 서열이 모체 유래의 비메틸화된 RASSF1A 서열을 배경으로 하여 임신 여성의 혈장에서 검출됨을 밝혀냈다 (문헌 [Chan et al, Clin Chem 2006;52:2211-8]). 따라서, 태아 DNA의 분획 농도는 메틸화된 RASSF1A 서열의 양을 전체 RASSF1A (메틸화 및 비메틸화 RASSF1A) 서열의 양으로 나눔으로써 측정될 수 있다.

[0304] 혈액 응고 동안 모체 혈액 세포로부터 DNA가 방출되기 때문에, 본 발명을 실행하기 위해서는 모체 혈장이 모체 혈청보다 바람직할 것으로 예상된다. 따라서, 혈청을 사용하는 경우, 태아 DNA의 분획 농도가 모체 혈청에서보다 모체 혈장에서 더 낮을 것으로 예상된다. 다르게 말하면, 모체 혈청을 사용하는 경우, 동일한 임신 여성으로부터 동시에 얻어지는 혈장 샘플과 비교했을 때, 진단하고자 하는 태아 염색체 이수성에 대해 더 많은 서열이 생성될 필요가 있을 것으로 예상된다.

[0305] 그러나, 태아 DNA의 분획 농도를 측정하기 위한 또다른 대안 방법은 임신 여성 및 태아 사이의 다형성 차이의 정량화를 통하는 것일 것이다 (문헌 [Dhallan R, et al. 2007 Lancet, 369, 474-481]). 이러한 방법의 예는 임신 여성이 동형접합성이고 태아가 이형접합성인 다형성 부위를 표적으로 할 것이다. 태아-특이적 대립유전자의 양을 공통 대립유전자의 양과 비교하여 태아 DNA의 분획 농도를 측정할 수 있다.

[0306] 비교 게놈 혼성화, 마이크로어레이 비교 게놈 혼성화, 정량적 실시간 중합효소 연쇄 반응 (하나 이상의 특이적 서열(들)을 검출하고 정량화함)을 비롯한, 염색체 이상을 검출하기 위한 기존의 기술과 달리, 대규모 병렬 서열분석은 미리 측정되거나 또는 미리 한정된 세트의 DNA 서열의 검출 또는 분석에 따라 좌우되지 않는다. 표본 풀로부터의 DNA 분자의 랜덤한 대표 분획을 서열분석한다. 다양한 염색체 영역에 대해 정렬된 상이한 서열 태그의 갯수는 종양성 DNA를 함유하거나 또는 함유하지 않은 표본 사이에서 비교한다. 염색체 이상은 표본에서 임의의 소정 염색체 영역에 대해 정렬된 서열의 갯수 (또는 백분율)의 차이로 나타낼 것이다.

[0307] 또다른 예에서, 혈장 무-세포 DNA에 대한 서열분석 기술을 이용하여 특정 암의 검출을 위해 혈장 DNA에서의 염색체 이상을 검출할 수 있다. 상이한 암은 한 세트의 전형적인 염색체 이상을 갖는다. 다중 염색체 영역의 변화 (증폭 및 결실)가 사용될 수 있다. 따라서, 증폭된 영역에 대해 정렬된 서열의 비율이 증가하고, 감소된 영역에 대해 정렬된 서열의 비율이 감소될 것이다. 염색체 당 표출 백분율은 전체 게놈과 관련하여 임의의 소정 염색체의 게놈 표출의 백분율로서 표현된 참조 게놈 중 각각의 상응하는 염색체에 대한 크기와 비교할 수 있다. 직접 비교 또는 참조 염색체에 대한 비교가 또한 사용될 수 있다.

[0309] VIII. 돌연변이 검출

[0310] 모체 혈장 중의 태아 DNA는 소량의 집단으로 존재하며, 모체 혈장 DNA 중 평균 3% 내지 6%가 태아에 의한다. 이러한 이유로 인해서, 본 분야에서 행해진 기존 연구의 대부분은 태아가 부계로부터 물려받고 모체 혈장 중의 대다수의 모체 DNA 배경과 구별되는 DNA 표적을 검출하는데 집중되었다. 이와 같이 기존에 검출된 표적의 예는 Y 염색체상의 SRY 유전자 [Lo YMD et al. 1998 Am J Hum Genet, 62, 768-775] 및 모계가 RhD-음성인 경우의 RHD 유전자 [Lo YMD et al. 1998 N Engl J Med, 339, 1734-1738]를 포함한다.

[0311] 태아 돌연변이 검출을 위해서 모체 혈장을 사용하는 기존의 전략은 부계가 보인자(carrier)인 상염색체 우성 상태로 제한되며, 부계 및 모계가 상이한 돌연변이를 갖는 경우에는 직접적인 돌연변이 검출을 수행하거나 또는 연관 분석 [Ding C. et al 2004 Proc Natl Acad Sci USA 101, 10762-10767]을 실시하여 상염색체 열성 질환은 제외한다. 이러한 기존 전략에는 유의한 한계가 있었다. 예를 들어, 남성 파트너와 여성 파트너가 동일한 돌연변이를 갖는 쌍에서는 모체 혈장 중에서의 직접적인 돌연변이 검출로는 의미있는 태아 진단을 실시할 수 없다.

[0312] 이러한 시나리오는 도 23에 예시되어 있다. 상기 시나리오에서는, 3종의 가능한 태아 유전자형, NN, NM 및 MM이 존재하는데, 여기서 N은 정상적인 대립유전자를 나타내고, M은 돌연변이 대립유전자를 나타낸다. 돌연변이 대립유전자의 예는 낭포성 섬유증, 베타-지중해빈혈, 알파-지중해빈혈, 겸상적혈구 빈혈, 척수 근위축증, 선천성 부신 과형성 등을 일으키는 것을 포함한다. 이러한 장애의 다른 예는 OMIM (Online Mendelian Inheritance in Man) [www.ncbi.nlm.nih.gov/sites/entrez?db=OMIM&itool=toolbar]에서 찾을 수 있다. 모체 혈장에서, 대부분의 DNA는 모계로부터 유래되며 NM일 것이다. 상기한 3종의 태아 유전자형 중 어느 것에서도, 모체 혈장 중에서의 독특한 검출을 허용하는 임의의 독특한 태아 대립유전자는 존재하지 않는다. 따라서, 동상적인 전략은 여기에 적용될 수 없다.

[0313] 본원에 기재한 실시양태는 이러한 시나리오를 취급할 수 있다. 모계 및 태아가 둘다 NM인 시나리오에서는 N 대립유전자 및 M 대립유전자가 대립유전자 균형을 이루고 있을 것이다. 그러나, 모계가 NM이고 태아가 NN인 경우에는 모체 혈장 중에서 N 대립유전자가 과다표출된 대립유전자 불균형이 일어날 것이다. 한편, 모계가 NM이고 태아가 MM인 경우에는 모체 혈장 중에서 M 대립유전자가 과다표출된 대립유전자 불균형이 일어날 것이다. 따라서, 태아 돌연변이 검출을 위한 귀무가설에서는 태아가 NM 유전자형인 경우에 대립유전자 불균형이 존재하지 않는다고 가정한다. 대립가설에서는, 대립유전자 불균형이 존재하고 N 또는 M 대립유전자 중 어느 것이 과다표출되었는지의 여부에 따라 태아 유전자형은 NN 또는 MM일 수 있다고 가정한다.

[0314] 대립유전자 불균형의 존재 또는 부재는 본원에 기재한 실시양태를 사용한 디지탈 PCR로 결정될 수 있다. 첫번째 시나리오에서, 특정 부피의 모체 혈장은 모계로부터의 50개 세포 및 태아로부터의 50개 세포의 100개 세포로부터 방출된 DNA를 함유한다. 따라서, 이러한 부피의 혈장 중에서 태아 DNA의 분획 농도는 50%이다. 모계가 유전자형 NM인 경우에는 50개의 N 대립유전자 및 50개의 M 대립유전자가 모계에 의한다. 태아가 유전자형 NM인 경우에는 50개의 N 대립유전자 및 50개의 M 대립유전자가 태아에 의한다. 따라서, N 대립유전자와 M 대립유전자 (각각 총 100개 카피) 사이에는 대립유전자 불균형이 존재하지 않는다. 한편, 태아가 NN 유전자형인 경우에는 이러한 부피의 혈장 중에 100개의 태아-유래의 N 대립유전자가 존재할 것이다. 따라서, 총 150개의 N 대립유전자와 50개의 M 대립유전자가 존재할 것이다. 즉, N과 M 사이에는 N이 M에 대하여 3:1의 비율로 과다표출된 대립유전자 불균형이 존재할 것이다.

[0315] 반대의 상황에서, 태아가 MM 유전자형인 경우에는 이러한 부피의 혈장 중에 100개의 태아-유래의 M 대립유전자가 존재할 것이다. 따라서, 150개의 M 대립유전자와 50개의 N 대립유전자가 존재할 것이다. 즉, N과 M 사이에는 M이 N에 대하여 3:1의 비율로 과다표출된 대립유전자 불균형이 존재할 것이다. 이러한 대립유전자 불균형은 디지탈 PCR로 결정될 수 있다. 양성 웰의 수가 더 적은 대립유전자가 참조 주형으로 간주된다. 디지탈 RNA-SNP 및 디지탈 RCD 분석과 유사하게, 디지탈 PCR 실험에서 대립유전자의 실제 분포는 푸아송 확률 밀도 함수에 의해 좌우된다. 따라서, 본 시나리오에서는 대립유전자 불균형의 이론적인 정도가 3:1이지만, 대립유전자 불균형의 예상 정도는 디지탈 PCR 분석 동안에 웰 당 평균 주형 농도에 따라 달라질 것이다. 따라서, SPRT 분석에서와 같은, 웰 당 평균 참조 주형 농도 (m_r)에 직접한 해석 분별값이 사례 분류에 사용될 필요가 있다.

[0316] 추가로, 결정할 필요가 있는 대립유전자 불균형의 정도는 태아 DNA의 분획 농도에 따라 달라진다. 상기한 예와는 반대로, 특정 부피의 모체 혈장이 모계로부터의 90개 세포 및 태아로부터의 10개 세포의 100개 세포로부터 방출된 DNA를 함유하는 경우를 가정한다. 따라서, 이러한 부피의 혈장 중에서 태아 DNA의 분획 농도는 10%이

다. 모계가 유전자형 NM인 경우에는 90개의 N 대립유전자 및 90개의 M 대립유전자가 모계에 의한다. 태아가 유전자형 NM인 경우에는 10개의 N 대립유전자 및 10개의 M 대립유전자가 태아에 의한다. 따라서, N 대립유전자와 M 대립유전자 (각각 총 100개 카피) 사이에는 대립유전자 불균형이 존재하지 않는다. 한편, 태아가 NN 유전자형인 경우에는 이러한 부피의 혈장 중에 20개의 태아-유래의 N 대립유전자가 존재할 것이다. 따라서, 총 110개의 N 대립유전자와 90개의 M 대립유전자가 존재할 것이다.

[0317] 즉, N과 M 사이에는 N이 M에 대하여 과다표출된 대립유전자 불균형이 존재할 것이다. 반대의 상황에서, 태아가 MM 유전자형인 경우에는 이러한 부피의 혈장 중에 20개의 태아-유래의 M 대립유전자가 존재할 것이다. 따라서, 110개의 M 대립유전자와 90개의 N 대립유전자가 존재할 것이다. 즉, N과 M 사이에는 M이 과다표출된 대립유전자 불균형이 존재할 것이다. 태아 DNA의 분획 농도가 10%인 경우에 대립유전자 불균형의 이론적인 정도는 110:90일 것이고, 이는 상기 예에서 나타낸 바와 같이 50% 태아 DNA가 존재하는 경우의 3:1 비율과 상이하다. 따라서, SPRT 분석에서와 같은, 태아 DNA의 분획 농도에 적절한 해석 분별값이 사례 분류에 사용될 필요가 있다.

[0318] 따라서, 혈장 DNA를 추출할 것이다. 혈장 샘플 중 모체 및 태아 DNA의 양은 예를 들어 기존에 수립된 실시간 PCR 검정 [Lo, et al. 1998 Am J Hum Genet 62, 768-775] 또는 당업자에게 공지된 다른 유형의 정량화 수단, 예를 들어 SNP 마커 [Dhallan R et al. 2007 Lancet, 369, 474-481] 및 태아 후성 마커 [Chan KCA et al. 2006 Clin Chem, 52, 2211-2218]를 이용하여 정량한다. 태아 DNA의 백분율(%)을 계산한다. 이후, 디지탈 PCR 분석 동안에 각 반응 웰이 평균 1개의 주형 분자 (N 대립유전자일 수도 있고, 또는 M 대립유전자일 수도 있음)를 함유하도록 정량된 혈장 DNA 샘플 (예를 들어 희석하거나 농축된 것)을 제조한다. 한쌍의 프라이머, 및 2개의 태크만 프로브 (1개는 N 대립유전자에 특이적이고, 다른 하나는 M 대립유전자에 특이적임)를 사용하여 디지탈 PCR 분석을 수행한다. 오직 M에만 양성인 웰의 수 및 오직 N에만 양성인 웰의 수를 계수한다. 이들 웰의 비율을 사용하여 대립유전자 불균형의 증거가 있는지 여부를 결정한다. 대립유전자 불균형의 통계적인 증거는 당업자에게 공지된 방법, 예컨대 SPRT를 사용하여 찾을 수 있다. 이러한 분석의 한 변형법에서는 M에만 양성이거나 M 및 N에 양성인 웰의 수를 계수하는 것도 가능하고, N에만 양성이거나 M 및 N에 양성인 웰의 수를 계수하는 것도 가능하며, 이들 계수치의 비율을 도출하는 것도 가능하다. 한편 더, 대립유전자 불균형의 통계적인 증거는 당업자에게 공지된 방법, 예컨대 SPRT를 사용하여 찾을 수 있다.

[0319] 디지탈 RMD (상대적 돌연변이 양(Relative Mutant Dosage))라고 불리는 태아 유전자 돌연변이의 양 결정을 여성/남성 (XX/XY) DNA 혼합물을 사용하여 검증하였다. 남성 및 여성으로부터의 혈액 세포 DNA를 남성 DNA와 각각 혼합하여, 도 24a에 나타낸 바와 같이, XY의 배경에 XX 또는 XY 유전자형을 각각 25% 및 50%의 분획 농도로 갖는 샘플을 수득한다.

[0320] 추가로, 혈액 세포 샘플을 또한 12명의 남성 및 12명의 여성 대상체로부터 수득하였다. 여성 혈액 세포 DNA (유전자형 XX)를 3배 과량의 남성 혈액 세포 DNA (유전자형 XY)와 각각 혼합하여, XY 유전자형을 갖는 DNA 75%의 배경에 XX 유전자형을 갖는 DNA 25%를 함유하는 12개의 DNA 혼합물을 생성하였고, 이 결과는 도 24b에 나타냈다.

[0321] SPRT 분석의 목적은 배경 DNA에 존재하는 소량의 유전자형을 결정하는 것이었다. 75% XY DNA의 배경에 25% XX DNA를 함유하는 DNA 혼합물에서, 소량의 대립유전자는 DNA 75%에서 유래된 Y이다. 샘플 중의 DNA 25%가 XX 유전자형이기 때문에, 샘플 중에 총 200개의 DNA 분자가 존재하는 경우에는 150개의 분자가 XY 개체에서 유래된 것이다. 따라서, Y 대립유전자의 수는 75개일 것으로 예상된다. 또한, 남성 DNA (유전자형 XY)의 비율에 의한 X 대립유전자의 수 역시 75개이다. 여성 (유전자형 XX)에 의한 X 대립유전자의 수는 50 (= 2×25)이다. 따라서, X/Y 비율은 125/75 = (1 + 25%)/(1 - 25%) = 5/3이다.

[0322] 본 연구의 두번째 부분에서, 혈액 세포 샘플을 베타-글로빈 유전자, 즉, 헤모글로빈, 베타 (HBB) 유전자에 HbE (G→A) 및 CD41/42 (CTTT/-) 돌연변이를 갖는 남성 및 여성 대상체로부터 수득하였다. 모든 가능한 유전자형 (MM, MN 또는 NN)을 갖는 남성 태아를 보유하는 이형접합 모계 (MN (여기서, M = 돌연변이, N = 야생형))로부터 수득한 모체 혈장 샘플을 모방하기 위해서, 야생형 대립유전자에 대해 동형접합 (NN)이거나 2종의 돌연변이 중 하나에 대해 이형접합 (MN)인 남성으로부터의 혈액 세포 DNA를 동일 돌연변이에 대하여 이형접합 (MN)인 여성으로부터 수집한 혈액 세포 DNA 샘플과 각각 혼합하였다. 이로써, 다양한 분획 농도의 남성/돌연변이 DNA를 갖는 DNA 혼합물이 생성되었다. 또한, CD41/42 결실에 대해 동형접합 (MM)인 여성으로부터의 혈액 세포 DNA 샘플도 사용하여 DNA 혼합물을 제조하였다. SPRT 분류를 위한 남성 비율을 정확하게 하기 위해서, 각 DNA 혼합물 중 남성 DNA의 분획 농도를 ZFY/X 검정으로 결정하였다.

[0323] 디지탈 ZFY/X 검정을 이용하여 SPRT를 검증하고 또한 DNA 혼합물 중 남성 DNA의 분획 농도를 결정하였다. 염색체 X에서의 아연 핑거(Zinc Finger) 단백질 서열 (ZFX) 및 염색체 Y에서의 아연 핑거 단백질 서열 (ZFY)의 양을 디지탈 PCR 분석으로 결정하였다. 전방향 프라이머 5'-CAAGTGCTGGACTCAGATGTAACTG-3' (서열 47) 및 역방향 프라이머 5'-TGAAGTAATGTCAGAAGCTAAAACATCA-3' (서열 48)를 사용하여 ZFX 및 ZFY 유전자좌의 87-bp 앰플리콘을 우선 동시 증폭시켰다. 염색체 X 및 Y 파라로그를 구별하는 2종의 염색체-특이적 테크만 프로브를 설계하였고, 이의 서열은 각각 5'-(VIC)TCTTTAGCAGATTGCA(MGBNFQ)-3' (서열 49) 및 5'-(FAM)TCTTTAGCACACTGCAC(MGBNFQ)-3' (서열 50)였다.

[0324] DNA 혼합물 중의 돌연변이 양을 돌연변이 대립유전자에 대한 정상 대립유전자의 디지탈 PCR 분석으로 결정하였다. HbE 돌연변이의 경우, 정상 및 돌연변이 대립유전자의 87-bp 앰플리콘을 전방향 프라이머 5'-CGGCAAGGTGAACGTGGAT-3' (서열 51) 및 역방향 프라이머 5'-CTATTGGTCTCCTTAAACCTGTCTTGTAA-3' (서열 52)로 우선 동시 증폭시켰다. 정상 (G) 및 돌연변이 (A) 대립유전자 사이를 구별하는 2종의 대립유전자-특이적 테크만 프로브를 설계하였고, 이의 서열은 각각 5'-(VIC)TTGGTGGTGAGGCC(MGBNFQ)-3' (서열 53) 및 5'-(FAM)TTGGTGGTAAGGCC(MGBNFQ)-3' (서열 54)였다. HbE 돌연변이에 대한 결과를 도 25에 나타냈다.

[0325] CD41/42 결실 돌연변이의 경우, 정상 및 돌연변이 대립유전자의 87-bp 및 83-bp 앰플리콘을 전방향 프라이머 5'-TTTTCCCACCCTTAGGCTGC-3' (서열 55) 및 역방향 프라이머 5'-ACAGCATCAGGAGTGGACAGATC-3' (서열 56) 각각을 사용하여 우선 동시 증폭시켰다. 정상 (결실 없음) 및 돌연변이 (결실 있음) 대립유전자 사이를 구별하는 2종의 대립유전자-특이적 테크만 프로브를 설계하였고, 이의 서열은 각각 5'-(VIC)CAGAGGTTCTTTGAGTCCT(MGBNFQ)-3' (서열 57) 및 5'-(FAM)AGAGGTTGACTCCTT(MGBNFQ)-3' (서열 58)였다. HbE 돌연변이에 대한 결과를 도 26a 및 도 26b에 나타냈다.

[0326] 이러한 실험은 12.765 디지탈 어레이즈 (플루이다임)를 사용하는 바이오마크™ 시스템 (플루이다임)에서 수행하였다. 1개의 패널에 대한 반응은 2× 태크만 유니버셜 PCR 마스타 믹스 (어플라이드 바이오시스템즈)를 사용하여 반응 부피 10 ㎕로 설정하였다. CD41/42 및 ZFY/X 검정의 경우, 각 반응물은 1× 태크만 유니버셜 PCR 마스타 믹스, 900 nM씩의 각 프라이머, 125 nM씩의 각 프로브 및 1 ng/㎕의 DNA 혼합물 3.5 ㎕를 함유하였다. HbE 검정의 경우, 정상 (G) 및 돌연변이 (A) 대립유전자를 표적으로 하는, 각각 250 nM 및 125 nM의 프로브를 첨가하였다. 샘플/검정 혼합물을 나노플렉스™ IFC 제어기 (플루이다임)로 디지탈 어레이에 로딩하였다. 반응은 신호 검출을 위해 바이오마크™ 시스템에서 수행하였다. 반응을 50℃에서 2분 동안 게시한 후에 95℃에서 10분 동안 진행하고, 95℃, 15초 및 57℃ (ZFY/X 및 CD41/42의 경우) 또는 56℃ (HbE의 경우), 1분의 50회 주기를 수행하였다. 각 경우마다 적어도 1개의 반응 패널을 사용하였고, 결정이 이루어질 수 있을 때까지 미분류 상태로 있는 샘플을 위한 여분의 패널로부터 데이타를 합하였다.

[0327] 또한, 디지탈 PCR이 당업자에게 공지된 방법, 예를 들어 미세유체 칩, 나노리터 PCR 마이크로플레이트 시스템, 에멀젼 PCR, 폴로니 PCR, 롤링-써클 증폭법, 프라이머 연장법 및 질량 분광법을 이용하여 수행될 수 있다는 것이 당업자에게 명백할 것이다.

[0329] IX. 암을 이용한 예

[0330] 한 실시양태에서, 본 발명을 수행하여 암성 종양에서 발생할 수 있는 것과 같이 대립유전자 비율이 벗어나거나 벗어나지 않은 샘플을 분류할 수 있다. 한 측면에서, 각 경우에 오직 A 대립유전자에 대해서만 양성 신호를 갖는 웰의 수, 오직 G 대립유전자에 대해서만 양성 신호를 갖는 웰의 수, 및 2종의 대립유전자 둘다에 대해서 양성 신호를 갖는 웰의 수를 디지탈 PCR로 결정하였다. 참조 대립유전자는 양성 웰의 수가 더 적은 대립유전자로 정의된다 (2종의 대립유전자가 둘다 동일한 수의 양성 웰을 갖는 시나리오가 있다고 한다면, 둘 중 어느 것이라도 참조 대립유전자로 사용될 수 있음). 참조 대립유전자에 대해 음성인 웰의 총 수를 사용하여, 다른 대립유전자가 양성인지 여부와는 관계 없이 푸아송 확률 밀도 함수에 따라 웰 당 참조 대립유전자의 평균 농도 추론치 (m_r)를 계산하였다. 본 연구진은 상기 계산을 예시하는 가정적인 예를 사용한다.

[0331] 96웰 반응에서, 20개 웰은 A 대립유전자 대해 양성이고, 24개 웰은 G 대립유전자에 대해 양성이며, 28개 웰은 이들 2개의 대립유전자 모두에 대해 양성이다. A 대립유전자에 대하여 양성인 웰의 수가 더 적기 때문에, A 대립유전자를 참조 대립유전자로 간주한다. 참조 대립유전자에 대해 음성인 웰의 수는 96 - 20 - 28 = 48개이다.

따라서, m_r은 푸아송 분포로 계산될 수 있고, -ln(48/96) = 0.693이다.

[0332] LOH 검출과 관련하여, 귀무가설에서는 1개의 대립유전자 결실이 존재하여 샘플에서 대립유전자 비율이 벗어나지 않았다고 가정한다. 이러한 가정하에서, 2종의 대립유전자에 대한 양성 웰의 수의 예상 비율은 1:1이고, 이에 따라 잠재적으로 과다표출된 대립유전자를 함유하는 정보제공 웰 (오직 1종의 대립유전자에 대해서만 양성인 웰)의 예상 비율은 0.5이다.

[0333] 299LOH 검출과 관련하여, 대립가설에서는 샘플의 세포 중 50%에서 1개의 대립유전자 결실이 존재하여 샘플에서 대립유전자 비율이 벗어나 있다고 가정한다. 과다표출된 대립유전자와 참조 대립유전자 사이의 대립유전자 비율이 2:1이기 때문에, 웰 당 과다표출된 대립유전자의 평균 농도는 참조 대립유전자의 평균 농도의 2배일 것이다. 그러나, 과다표출된 대립유전자에 양성인 웰의 수는 참조 대립유전자에 양성인 웰의 수의 단순히 2배가 아니라 푸아송 분포에 따를 것이다.

[0334] 정보제공 웰은 A 또는 G 대립유전자에 양성이지만 상기 2종의 대립유전자 둘다에는 양성이 아닌 웰로 정의된다. 대립유전자 비율이 벗어난 샘플의 경우에 과다표출된 대립유전자를 함유하는 웰의 수의 예상 비율의 계산치는 표 600에 나타낸 바와 같다. 상기 예에서, 종양 세포의 50%에서 LOH가 존재하는 경우에는 웰 당 G 대립유전자의 평균 농도가 $2\times0.693 = 1.386$일 것이다. LOH가 종양 세포의 50% 초과에 존재하는 경우에는 웰 당 G 대립유전자의 평균 농도가 하기 식에 따를 것이다: $1/[1 - (\text{LOH를 갖는 비율})] \times m_r$.

[0335] G 대립유전자에 양성인 웰의 예상 비율은 $1 - e^{-1.386} = 0.75$ (즉, 75% 또는 72개의 웰)일 것이다. A 또는 G 대립유전자에 대한 웰의 양성이 독립적이라고 가정할 때, $0.5 \times 0.75 = 0.375$의 웰이 A 및 G 대립유전자 둘다에 양성일 것이다. 따라서, $0.5 - 0.375 = 0.125$의 웰이 오직 A 대립유전자에만 양성일 것이고, $0.75 - 0.375 = 0.375$의 웰이 오직 G 대립유전자에만 양성일 것일 것이다. 따라서, 정보제공 웰의 비율은 $0.125 + 0.375 = 0.5$일 것이다. G 대립유전자를 갖는 정보제공 웰의 예상 비율은 $0.375/0.5 = 0.75$이다. 이후, P_r에 대한 이러한 예상 값을 사용하여, 샘플에 대립유전자 비율이 벗어난 경우 (즉, 여기서는 LOH)가 존재하는지의 여부를 결정하기 위한 적절한 SPRT 곡선을 구축할 수 있다.

[0336] 이어서, 디지탈 PCR 분석을 통해 실험적으로 결정된, 비-참조 대립유전자를 갖는 정보제공 웰의 실제 비율 (P_r)을 사용하여, 귀무가설 또는 대립가설 중 어느 것이 받아들여질 지, 또는 더 많은 웰을 사용한 추가의 분석이 필요할 지를 결정하였다. 귀무가설 또는 대립가설을 받아들이기 위한 P_r의 결정 한계값은 임계값 우도 비율 8을 기준으로 하여 계산되는데, 이는 상기 값이 대립유전자 불균형이 존재하는 샘플과 존재하지 않는 샘플을 암 검출과 관련하여 식별하는데 만족스러운 성능을 제공하는 것으로 나타났기 때문이다 ([Zhou, W, et al. (2001) Nat Biotechnol 19, 78-81], [Zhou et al 2002 (상기 문헌)]). 상기 예에서, 정보제공 웰의 수는 20 + 24 = 44이고, 실험적으로 수득된 P_r은 24/44 = 0.5455이다. 결정 한계값은 귀무가설을 받아들이기 위해서는 $\leq$ 0.5879이고, 대립가설을 받아들이기 위해서는 $\geq$ 0.6739이다. 따라서, 이러한 예에서의 샘플은 대립유전자 비율이 벗어난 NOT으로 분류된다.

[0337] 결론적으로, 본 연구진은 샘플 중의 서열 불균형을 검출하는 접근법을 알아냈다. 한 실시양태에서, 본 발명은 태아 염색체 이수성의 비-침습적 검출에 사용될 수 있고, 예를 들어 모체 혈상 중 태아 핵산의 분석을 통해 삼염색체 21을 검출할 수 있다. 이러한 접근법은 또한 양수, 융모막 융모 샘플, 모체 뇨, 자궁경관 샘플, 모체 타액 등을 포함하는, 태아 핵산을 함유하는 다른 생물학적 물질에도 적용될 수 있다. 첫째, 본 연구진은 삼염색체 21 태아를 보유하는 여성의 모체 혈장 중 염색체 21에서의 태반-발현된 전사체인 PLAC4 mRNA에서 SNP의 대립유전자 불균형을 결정하기 위한 본 발명의 용도를 입증하였다. 둘째, 본 연구진은 본 발명이 삼염색체 21의 비-침습적 태아 검출을 위해서 RCD (상대적 염색체 양(Relative Chromosome Dosage)) 분석을 통해 비-다형성 기재의 방법으로 사용될 수 있음을 입증하였다. 이러한 디지탈 RCD-기반의 접근법은 태아 DNA를 함유하는 샘플 중 염색체 21의 전체 카피수가 참조 염색체에 비하여 과다표출되었는지 여부를 직접 평가하는 것을 수반한다. 디지탈 RCD는 정교한 장치 없이도 25% 태아 DNA를 함유하는 샘플에서 삼염색체 21을 검출할 수 있게 한다. 본 연구진은 순차 확률비 검증 (SPRT)을 적용하여 디지탈 PCR 데이타를 해석하였다. 컴퓨터 시뮬레이션 분석을 실시하여 질환 분류 알고리즘의 높은 정확성을 확인하였다.

[0338] 추가로, 본 연구진은 상기 접근법이 염색체 이수성 이외의 다른 형태의 핵산 서열 불균형의 결정, 예를 들어 모체 혈장 중 태아 돌연변이의 검출 또는 다형성 검출 및 혈장 중 종양-유래의 핵산 분석에 의한 악성 세포 게놈의 국소적인 획득 및 소실의 검출에 적용될 수 있음을 알아냈다.

[0339] 본 명세서에 기재한 임의의 소프트웨어 성분 또는 기능은 임의의 적합한 컴퓨터 언어, 예를 들어 자바(Java), C++ 또는 펄(Perl) (예를 들어 통상의 기술 또는 객체 지향(object-oriented) 기술을 사용한 것)을 사용한 프로세서에 의해 실행되는 소프트웨어 코드로서 시행될 수 있다. 소프트웨어 코드는 일련의 지시 또는 명령어로서 저장을 위한 컴퓨터 판독가능 매체에 저장될 수 있으며, 적합한 매체는 랜덤 액세스 메모리 (RAM, random access memory), 리드 온리 메모리 (ROM, read only memory), 자성 매체, 예컨대 하드 드라이브 또는 플로피 디스크, 또는 광학 매체, 예컨대 콤팩트 디스크 (CD) 또는 DVD (디지탈 버서틀 디스크(digital versatile disk)), 플래쉬 메모리 등을 포함한다. 컴퓨터 판독가능 매체는 이러한 저장 장치의 임의의 조합일 수 있다.

[0340] 이러한 프로그램은 또한 다양한 프로토콜, 예를 들어 인터넷에 따라 유선, 광학 및/또는 무선 네트워크를 통해 전송되도록 맞춰진 캐리어 신호를 사용하여 암호화되고 전송될 수도 있다. 이와 같이, 본 발명의 실시양태에 따른 컴퓨터 판독가능 매체가 이러한 프로그램으로 암호화된 데이타 신호를 사용하여 생성될 수 있다. 프로그램 코드로 암호화된 컴퓨터 판독가능 매체는 호환가능한 장치와 함께 패키지될 수도 있고, 또는 다른 장치와 별도로 제공 (예를 들어, 인터넷 다운로드에 의함)될 수도 있다. 임의의 이러한 컴퓨터 판독가능 매체는 1개의 컴퓨터 프로그램 제품 (예를 들어 하드 드라이브 또는 전체 컴퓨터 시스템)상에 있거나 그 내부에 있을 수 있고, 시스템 또는 네트워크 내부의 상이한 컴퓨터 프로그램 제품상에 있거나 그 내부에 있을 수 있다. 컴퓨터 시스템은 모니터, 프린터 또는 본원에서 언급한 임의의 결과를 사용자에게 제공하기 위한 다른 적합한 디스플레이를 포함할 수 있다.

[0341] 컴퓨터 시스템의 예는 도 27에 나타냈다. 도 27에 나타낸 하위시스템은 시스템 버스 (2775)를 통해 상호연결된다. 디스플레이 어댑터 (2782)에 커플링된 추가의 하위시스템, 예컨대 프린터 (2774), 키보드 (2778), 고정 디스크 (2779), 모니터 (2776) 등을 나타냈다. I/O 제어기 (2771)에 커플링된 주변기기 및 입력/출력 (I/O) 장치는 직렬 포트 (2777)과 같이 당업계에 공지된 임의의 수의 수단을 통해 컴퓨터 시스템에 연결될 수 있다. 예를 들어, 직렬 포트 (2777) 또는 외부 인터페이스 (2781)을 사용하여 컴퓨터 장치를 광범위한 범위의 네트워크, 예를 들어 인터넷, 마우스 입력 장치 또는 스캐너에 연결할 수 있다. 시스템 버스를 통한 상호연결은 중앙 프로세서 (2773)가 각각의 하위시스템과 소통하고 시스템 메모리 (2772) 또는 고정 디스크 (2779)로부터의 지시의 수행 및 하위시스템들 사이의 정보 교환을 제어하도록 한다. 시스템 메모리 (2772) 및/또는 고정 디스크 (2779)는 컴퓨터 판독가능 매체를 포함할 수 있다.

[0342] 본 발명의 예시적인 실시양태에 관한 상기 기재는 예시와 설명을 위해서 제공된 것이다. 이는 본 발명을 상기한 형태로만 한정하거나 제한하려는 것이 아니며, 상기한 기술에 비추어 많은 변형과 변화가 가능하다. 본 발명의 이론 및 이것의 실제적인 적용을 최적으로 설명하고 그에 따라 당업자가 본 발명을 고려되는 특별한 용도에 맞게 다양하게 변형시켜 다양한 실시양태에서 최적으로 사용할 수 있도록 하는 실시양태를 선택하여 기재하였다.

[0343] 본원에서 언급한 모든 간행물, 특허 및 특허 출원서는 어떠한 목적으로도 그 전문이 본원에 참고로 포함된다.

SEQUENCE LISTING

```
<110> Lo, Yuk-Ming Dennis
      Chiu, Rossa Wai Kwun
      Chan, Kwan Chee
      Zee, Benny Chung Ying
      Ra, Chun Chong
      The Chinese University of Hong Kong

<120> Determining Nucleic Acid Sequence
  Imbalance

<130> 016266-005219PC

<140> NO PCT/GB08/02524
<141> 2008-07-23

<150> US 60/961,438
<151> 2007-07-23

<160> 58

<170> FastSEQ for Windows Version 4.0

<210> 1
<211> 27
<212> DNA
<213> Artificial Sequence

<220>
<223> synthetic real-time PCR and digital RNA-DNP analysis
      PLAC4 gene-specific reverse transcription primer

<400> 1
agtstatagg accstgttta ggccagg                                    27

<210> 2
<211> 21
<212> DNA
<213> Artificial Sequence

<220>
<223> synthetic real-time PCR PLAC4 primer

<400> 2
ccgctaggct gtctttaag c                                           21

<210> 3
<211> 25
<212> DNA
<213> Artificial Sequence

<220>
<223> synthetic real-time PCR PLAC4 primer

<400> 3
ctgttacact acaaaataag tttct                                      25
```

[0344]

<210> 5
<211> 14
<212> DNA
<213> Artificial Sequence

<220>
<223> synthetic real-time PCR PLAC4 fluorescent probe

<220>
<221> modified_base
<222> (1)...(1)
<223> c modified by 6-carboxyfluorescein (FAM)

<220>
<221> modified_base
<222> (14)...(14)
<223> c modified by minor groove binding non-fluorescent
quencher (MGBNFQ)

<400> 4
attggagcaa sttc 14

<210> 5
<211> 104
<212> DNA
<213> Artificial Sequence

<220>
<223> single-stranded synthetic DNA oligonucleotide
amplicon

<400> 5
cgccgctagg gtgtctttta agctattgga gcaaattcaa atttggctta aagaaaaagc 60
aactcatttt gtattgcaac accaggagta tcccaaggga cttg 104

<210> 6
<211> 22
<212> DNA
<213> Artificial Sequence

<220>
<223> synthetic digital RNA-SNP analysis non-intron
spanning forward primer

<400> 6
tttgtattgc aacaccattt gg 22

<210> 7
<211> 15
<212> DNA
<213> Artificial Sequence

<220>
<223> synthetic allele-G-specific TaqMan probe targeting
rs[illegible] SNP of PLAC4

<220>
<221> modified_base
<222> (1)...(1)
<223> t modified by 6-carboxyfluorescein (FAM)

[0345]

<220>
<221> modified_base
<222> (15)...(15)
<223> g modified by minor groove binding non-fluorescent quencher (MGBNFQ)

<400> 7
tcgtcgtcta actcg 15

<210> 8
<211> 17
<212> DNA
<213> Artificial Sequence

<220>
<223> synthetic allele-A-specific TaqMan probe targeting rs8130833 SNP of FLAC4

<220>
<221> modified_base
<222> (1)...(1)
<223> a modified by fluorescent reporter VIC

<220>
<221> modified_base
<222> (17)...(17)
<223> g modified by minor groove binding non-fluorescent quencher (MGBNFQ)

<400> 8
attcgtcatc taactcg 17

<210> 9
<211> 21
<212> DNA
<213> Artificial Sequence

<220>
<223> synthetic digital RCD analysis co-amplification forward primer for paralogous loci on chromosome 21 and 1

<400> 9
gttgttctgc aaaaaacctt cga 23

<210> 10
<211> 28
<212> DNA
<213> Artificial Sequence

<220>
<223> synthetic digital RCD analysis co-amplification reverse primer for paralogous loci on chromosome 21 and 1

<400> 10
cttgcccaga aatacttcat tcccatat 28

[0346]

<210> 11
<211> 19
<212> DNA
<213> Artificial Sequence

<220>
<223> synthetic chromosome-specific TaqMan probe
designed to target chromosome 21 paralog

<220>
<221> modified_base
<222> (1)...(1)
<223> t modified by 6-carboxyfluorescein (FAM)

<220>
<221> modified_base
<222> (19)...(19)
<223> a modified by minor groove binding non-fluorescent
quencher (MGBNFQ)

<400> 11
taccccate atgagtaaa 19

<210> 12
<211> 20
<212> DNA
<213> Artificial Sequence

<220>
<223> synthetic chromosome-specific TaqMan probe
designed to target chromosome 1 paralog

<220>
<221> modified_base
<222> (1)...(1)
<223> c modified by fluorescent reporter VIC

<220>
<221> modified_base
<222> (20)...(20)
<223> a modified by minor groove binding non-fluorescent
quencher (MGBNFQ)

<400> 12
cgtaccctctg taatgtgtaa 20

<210> 13
<211> 27
<212> DNA
<213> Artificial Sequence

<220>
<223> synthetic digital PCR PLAC4 gene-specific reverse
transcription primer

<400> 13
agtatatagg accatgtttg ggccaga 27

4

[0347]

<210> 14
<211> 22
<212> DNA
<213> Artificial Sequence

<220>
<223> synthetic digital PCR amplification forward primer
for rs61130833 SNP region of PLAC4 gene

<400> 14
ccgtaattgc aatacaattt gg 22

<210> 15
<211> 13
<212> DNA
<213> Artificial Sequence

<220>
<223> synthetic allele-G-specific TaqMan probe targeting
PLAC4 rs61130833 SNP

<220>
<221> modified_base
<222> (1)...(1)
<223> t modified by 5-carboxyfluorescein (FAM)

<220>
<221> modified_base
<222> (13)...(13)
<223> g modified by minor groove binding non-fluorescent
quencher (MGBNFQ)

<400> 15
tcgtagtcct aattg 15

<210> 16
<211> 17
<212> DNA
<213> Artificial Sequence

<220>
<223> synthetic allele-A-specific TaqMan probe targeting
PLAC4 rs61130833 SNP

<220>
<221> modified_base
<222> (1)...(1)
<223> c modified by fluorescent reporter VIC

<220>
<221> modified_base
<222> (17)...(17)
<223> g modified by minor groove binding non-fluorescent
quencher (MGBNFQ)

<400> 16
atcgtcctc taattg 17

5

[0348]

<210> 17
<211> 20
<212> DNA
<213> Artificial Sequence

<220>
<223> synthetic digital PCR amplification serpin
peptidase inhibitor clade B (ovalbumin) member 2
(SERPINB2) gene-specific reverse transcription
primer

<400> 17
cccagacttc tcaccaacca 20

<210> 18
<211> 20
<212> DNA
<213> Artificial Sequence

<220>
<223> synthetic digital PCR amplification serpin
peptidase inhibitor clade B (ovalbumin) member 2
(SERPINB2) forward primer

<400> 18
ccagcctctg caatcaatgc 20

<210> 19
<211> 16
<212> DNA
<213> Artificial Sequence

<220>
<223> synthetic allele-A-specific TaqMan probe targeting
serpin peptidase inhibitor clade B (ovalbumin)
member 2 (SERPINB2) rs6098 SNP

<220>
<221> modified base
<222> (1)...(1)
<223> c modified by 6-carboxyfluorescein (FAM)

<220>
<221> modified base
<222> (16)...(16)
<223> t modified by minor groove binding non-fluorescent
quencher (NGBRFQ)

<400> 19
ccacagggaa ttattt 16

<210> 20
<211> 16
<212> DNA
<213> Artificial Sequence

<220>
<223> synthetic allele-G-specific TaqMan probe targeting
serpin peptidase inhibitor clade B (ovalbumin)
member 2 (SERPINB2) rs6098 SNP

6

[0349]

<220>
<221> modified_base
<222> (1)...(1)
<223> c modified by 6-carboxyfluorescein (FAM)

<220>
<221> modified_base
<222> (16)...(16)
<223> t modified by minor groove binding non-fluorescent
quencher (MGBNFQ)

<400> 20
ccscagggga ttcttt 16

<210> 21
<211> 23
<212> DNA
<213> Artificial Sequence

<220>
<223> synthetic digital PCR analysis co-amplification
forward primer for 101-bp amplicon for paralogous
loci on chromosome 21 and 1

<400> 21
gttgttctgc aaaaaacctt cgg 23

<210> 22
<211> 28
<212> DNA
<213> Artificial Sequence

<220>
<223> synthetic digital PCR analysis co-amplification
reverse primer for 101-bp amplicon for paralogous
loci on chromosome 21 and 1

<400> 22
cttggccaga aatacttcat tacccatat 28

<210> 23
<211> 19
<212> DNA
<213> Artificial Sequence

<220>
<223> synthetic chromosome-specific TaqMan probe
designed to target chromosome 21 paralog

<220>
<221> modified_base
<222> (1)...(1)
<223> t modified by 6-carboxyfluorescein (FAM)

<220>
<221> modified_base
<222> (19)...(19)
<223> a modified by minor groove binding non-fluorescent
quencher (MGBNFQ)

[0350]

<400> 23
tacctcctc atgcgtcac 19

<210> 24
<211> 25
<212> DNA
<213> Artificial Sequence

<220>
<223> synthetic chromosome-specific TaqMan probe
designed to target chromosome 1 paralog

<220>
<221> modified_base
<222> (1)...(1)
<223> a modified by fluorescent reporter VIC

<220>
<221> modified_base
<222> (20)...(20)
<223> a modified by minor groove binding non-fluorescent
quencher (MGBNFQ)

<400> 24
cgtacctctg tastgtgtaa 25

<210> 25
<211> 24
<212> DNA
<213> Artificial Sequence

<220>
<223> synthetic digital RCD analysis co-amplification
forward primer for 128-bp amplicon for paralogous
loci on chromosome 21 and 18

<400> 25
gtcagaaac cacaaactga tcgg 24

<210> 26
<211> 17
<212> DNA
<213> Artificial Sequence

<220>
<223> synthetic digital RCD analysis co-amplification
reverse primer for 128-bp amplicon for paralogous
loci on chromosome 21 and 18

<400> 26
gtccaggctg tgggcct 17

<210> 27
<211> 18
<212> DNA
<213> Artificial Sequence

<220>
<223> synthetic chromosome-specific TaqMan probe
designed to target chromosome 21 paralog

[0351]

<220>
<221> modified_base
<222> (1)...(1)
<223> a modified by 6-carboxyfluorescein (FAM)

<220>
<221> modified_base
<222> (14)...(14)
<223> a modified by minor groove binding non-fluorescent quencher (MGBNFQ)

<400> 27
aagaggcgag gcaa 14

<210> 28
<211> 15
<212> DNA
<213> Artificial Sequence

<220>
<223> synthetic chromosome-specific TaqMan probe designed to target chromosome 18 paralog

<220>
<221> modified_base
<222> (1)...(1)
<223> a modified by fluorescent reporter VIC

<220>
<221> modified_base
<222> (15)...(15)
<223> a modified by minor groove binding non-fluorescent quencher (MGBNFQ)

<400> 28
aaaaaaaag gcaac 15

<210> 29
<211> 33
<212> DNA
<213> Artificial Sequence

<220>
<223> synthetic digital PCR analysis co-amplification forward primer for 121-bp amplicon for paralogous loci on chromosome 21 and 1

<400> 29
acgttggatg gttgttctgc aaaaaacctt cga 33

<210> 30
<211> 38
<212> DNA
<213> Artificial Sequence

<220>
<223> synthetic digital PCR analysis co-amplification reverse primer for 121-bp amplicon for paralogous loci on chromosome 21 and 1

[0352] 9

<400> 30
acgttggatg cttggcaaga aatacttcat taccatat 38

<210> 31
<211> 21
<212> DNA
<213> Artificial Sequence

<220>
<223> synthetic digital PCR analysis co-amplification
extension primer for 121-bp amplicon for
paralogous loci on chromosome 21 and 1

<400> 31
ctcatcctca cttcgtaccc c 21

<210> 32
<211> 34
<212> DNA
<213> Artificial Sequence

<220>
<223> synthetic digital RCD analysis co-amplification
forward primer for 148-bp amplicon for paralogous
loci on chromosome 21 and 18

<400> 32
acgttggatg gtacagaaac cacaaccgc tcgg 34

<210> 33
<211> 27
<212> DNA
<213> Artificial Sequence

<220>
<223> synthetic digital RCD analysis co-amplification
reverse primer for 148-bp amplicon for paralogous
loci on chromosome 21 and 18

<400> 33
acgttggatg gtccaggctg tgggcct 27

<210> 34
<211> 17
<212> DNA
<213> Artificial Sequence

<220>
<223> synthetic digital RCD analysis co-amplification
extension primer for 148-bp amplicon for
paralogous loci on chromosome 21 and 18

<400> 34
acaaaaggcg gaagagg 17

<210> 35
<211> 35
<212> DNA
<213> Artificial Sequence

[0353]

<220>
<223> synthetic digital RCD analysis co-amplification
forward primer for 81 bp amplicon for paralogous
loci on chromosome 21 and 1

<400> 35
acgttggatg ttgatgaagt ctcctctcta cttcg 35

<210> 36
<211> 33
<212> DNA
<213> Artificial Sequence

<220>
<223> synthetic digital RCD analysis co-amplification
reverse primer for 81 bp amplicon for paralogous
loci on chromosome 21 and 1

<400> 36
acgttggatg caatcagccc ggccagaaat act 33

<210> 37
<211> 37
<212> DNA
<213> Artificial Sequence

<220>
<223> synthetic digital RCD analysis co-amplification
forward primer for 82 bp amplicon for paralogous
loci on chromosome 21 and 7

<400> 37
acgttggatg gaatttaagc taaatcagcc tgaactg 37

<210> 38
<211> 37
<212> DNA
<213> Artificial Sequence

<220>
<223> synthetic digital RCD analysis co-amplification
reverse primer for 82 bp amplicon for paralogous
loci on chromosome 21 and 7

<400> 38
acgttggatg gtttctcata gttcatagta ggcttat 37

<210> 39
<211> 30
<212> DNA
<213> Artificial Sequence

<220>
<223> synthetic digital RCD analysis co-amplification
forward primer for 101 bp amplicon for paralogous
loci on chromosome 21 and 2

<400> 39
acgttggatg ccaggcaggg ttctatgcag 30

[0354] 11

<210> 40
<211> 30
<212> DNA
<213> Artificial Sequence

<220>
<223> synthetic digital RCD analysis co-amplification
reverse primer for 101 bp amplicon for paralogous
loci on chromosome 21 and 2

<400> 40
acgttggatg aggcggctic ctggctcttg 30

<210> 41
<211> 31
<212> DNA
<213> Artificial Sequence

<220>
<223> synthetic digital RCD analysis co-amplification
forward primer for 102 bp amplicon for paralogous
loci on chromosome 21 and 6

<400> 41
acgttggatg gctcgtctca ggctcgtagt t 31

<210> 42
<211> 30
<212> DNA
<213> Artificial Sequence

<220>
<223> synthetic digital RCD analysis co-amplification
reverse primer for 102 bp amplicon for paralogous
loci on chromosome 21 and 6

<400> 42
acgttggatg ttcttcgag cccttcttgg 30

<210> 43
<211> 23
<212> DNA
<213> Artificial Sequence

<220>
<223> synthetic extension primer for paralogous sequence
mismatches (PSM) on paralogous loci on chromosomes
21 and 1

<400> 43
gtctcatctc tacttcgtac ctc 23

<210> 44
<211> 19
<212> DNA
<213> Artificial Sequence

<220>
<223> synthetic extension primer for paralogous sequence
mismatches (PSM) on paralogous loci on chromosomes
21 and 7

12

[0355]

<400> 44
ttttacgctg tccccattt 19

<210> 45
<211> 19
<212> DNA
<213> Artificial Sequence

<220>
<223> synthetic extension primer for paralogous sequence mismatches (PSM) on paralogous loci on chromosomes 21 and 2

<400> 45
ggtctatgca ggagcagac 19

<210> 46
<211> 23
<212> DNA
<213> Artificial Sequence

<220>
<223> synthetic extension primer for paralogous sequence mismatches (PSM) on paralogous loci on chromosomes 21 and 6

<400> 46
tgggcgcggg agcggacttc gctgg 25

<210> 47
<211> 25
<212> DNA
<213> Artificial Sequence

<220>
<223> synthetic digital PCR analysis co-amplification forward primer for 87-bp amplicon for zinc finger protein sequences on chromosome X (ZFX) and Y (ZFY)

<400> 47
caagtgctgg actcagatgt aactg 25

<210> 48
<211> 29
<212> DNA
<213> Artificial Sequence

<220>
<223> synthetic digital PCR analysis co-amplification reverse primer for 87-bp amplicon for zinc finger protein sequences on chromosome X (ZFX) and Y (ZFY)

<400> 48
tgaagtaatg tcagaagcta aaacatca 28

<210> 49
<211> 16
<212> DNA
<213> Artificial Sequence

13

[0356]

<220>
<223> synthetic chromosome-specific TaqMan probe
designed to target Chromosome X paralog

<220>
<221> modified_base
<222> (1)...(1)
<223> c modified by fluorescent reporter VIC

<220>
<221> modified_base
<222> (15)...(15)
<223> a modified by minor groove binding non-fluorescent
quencher (MGBNFQ)

<400> 49
tcttcagcac attgca 16

<210> 50
<211> 17
<212> DNA
<213> Artificial Sequence

<220>
<223> synthetic chromosome-specific TaqMan probe
designed to target chromosome Y paralog

<220>
<221> modified_base
<222> (1)...(1)
<223> c modified by 6-carboxyfluorescein (FAM)

<220>
<221> modified_base
<222> (17)...(17)
<223> c modified by minor groove binding non-fluorescent
quencher (MGBNFQ)

<400> 50
tcttcacaac actgcac 17

<210> 51
<211> 19
<212> DNA
<213> Artificial Sequence

<220>
<223> synthetic digital PCR co-amplification forward
primer for 87-bp amplicon of beta-globin
(hemoglobin beta [HBB]) HbS mutation forward
primer

<400> 51
gggcaaggtg acaggtgaa 19

<210> 52
<211> 29
<212> DNA
<213> Artificial Sequence

[0357] 표 11

<220>
<223> synthetic digital PCR co-amplification forward primer for 87-bp amplicon of beta-globin (hemoglobin beta (HBB)) HbE mutation reverse primer

<400> 52
ctattggtct ccttaaacct gtcttgtaa 29

<210> 53
<211> 14
<212> DNA
<213> Artificial Sequence

<220>
<223> synthetic allele-G-specific TaqMan probe targeting beta-globin (hemoglobin beta (HBB)) HbE mutation normal allele

<220>
<221> modified_base
<222> (1)...(1)
<223> t modified by fluorescent reporter VIC

<220>
<221> modified_base
<222> (14)...(14)
<223> c modified by minor groove binding non-fluorescent quencher (MGBNFQ)

<400> 53
ttggtggtga ggcc 14

<210> 54
<211> 14
<212> DNA
<213> Artificial Sequence

<220>
<223> synthetic allele-A-specific TaqMan probe targeting beta-globin (hemoglobin beta (HBB)) HbE mutation mutant allele

<220>
<221> modified_base
<222> (1)...(1)
<223> t modified by 6-carboxyfluorescein (FAM)

<220>
<221> modified_base
<222> (14)...(14)
<223> c modified by minor groove binding non-fluorescent quencher (MGBNFQ)

<400> 54
ttggtgatga ggcc 14

<210> 55
<211> 20
<212> DNA
<213> Artificial Sequence

15

[0358]

<220>
<223> synthetic digital PCR co-amplification forward primer for 87-bp and 83-bp amplicon of CD41/42 deletion mutation normal and mutant allele

<400> 55
ttttccccc cttaggctgc 20

<210> 56
<211> 23
<212> DNA
<213> Artificial Sequence

<220>
<223> synthetic digital PCR co-amplification reverse primer for 87-bp and 83-bp amplicon of CD41/42 deletion mutation normal and mutant allele

<400> 56
acccccccag gagtggccag atc 23

<210> 57
<211> 19
<212> DNA
<213> Artificial Sequence

<220>
<223> synthetic allele-specific TaqMan probe targeting CD41/42 deletion mutation normal allele (without deletion)

<220>
<221> modified_base
<222> (1)...(1)
<223> c modified by fluorescent reporter VIC

<220>
<221> modified_base
<222> (19)...(19)
<223> t modified by minor groove binding non-fluorescent quencher (MGBNFQ)

<400> 57
cagaggttct ttgagtcct 19

<210> 58
<211> 15
<212> DNA
<213> Artificial Sequence

<220>
<223> synthetic allele-specific TaqMan probe targeting CD41/42 deletion mutation mutant allele (with deletion)

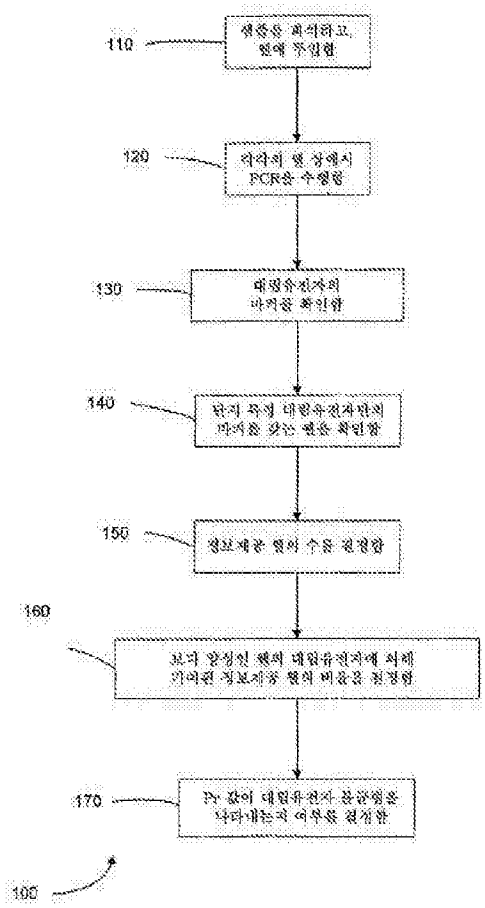
<220>
<221> modified_base
<222> (1)...(1)
<223> a modified by 6-carboxyfluorescein (FAM)

[0359]

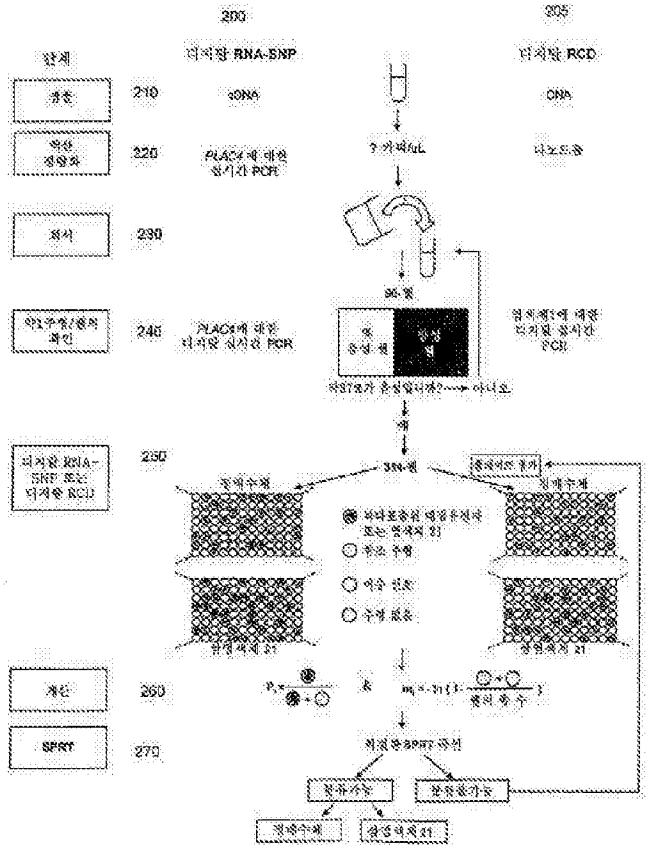
16

도면

도면1



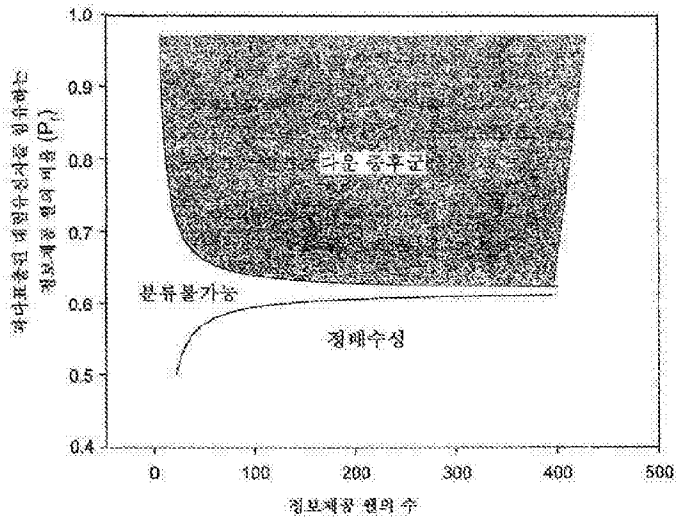
도면2a



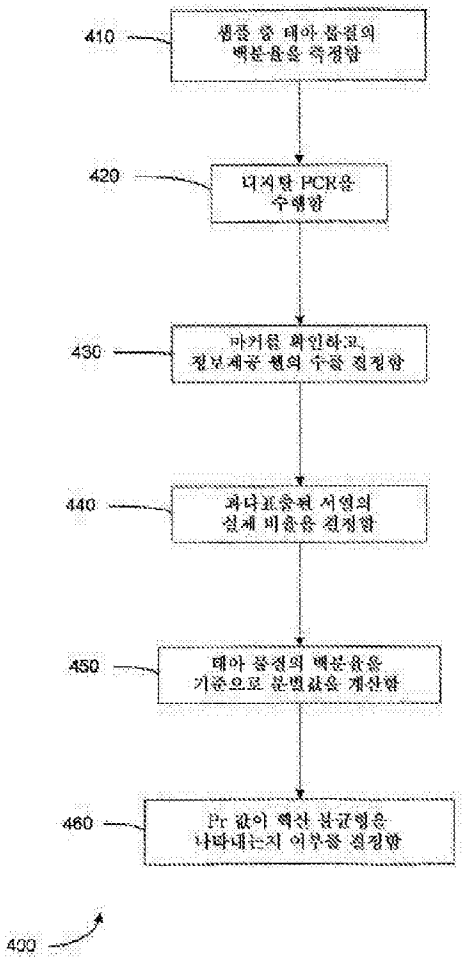
도면2b

암	동상적인 전액체 이상	
	희석	소실
방광암	7	9p; 11q; 17p
유방암	1q; 8q; 17q; 20q	8p; 13q; 16q; 18q
전장기종양	7; 8q; 13q; 20q	1p; 5q; 8p; 15q; 17p; 18q
전세포 양성	1q; 8q; 20q	4q; 8p; 13q; 16q; 17p
비소세포 폐암종	1q; 3q; 5p; 8q	3p; 8p; 9p; 13q; 17p
소세포 폐암종	3q; 5p; 8q; 19q	3p; 4q; 5q; 10q; 13q; 16q
미만두 암종	1q; 3q; 8; 12; 19	1p; 3p; 9; 11q; 13q; 14q; 16q
전립선암	7; 8q; 20q	8p; 13q; 16q

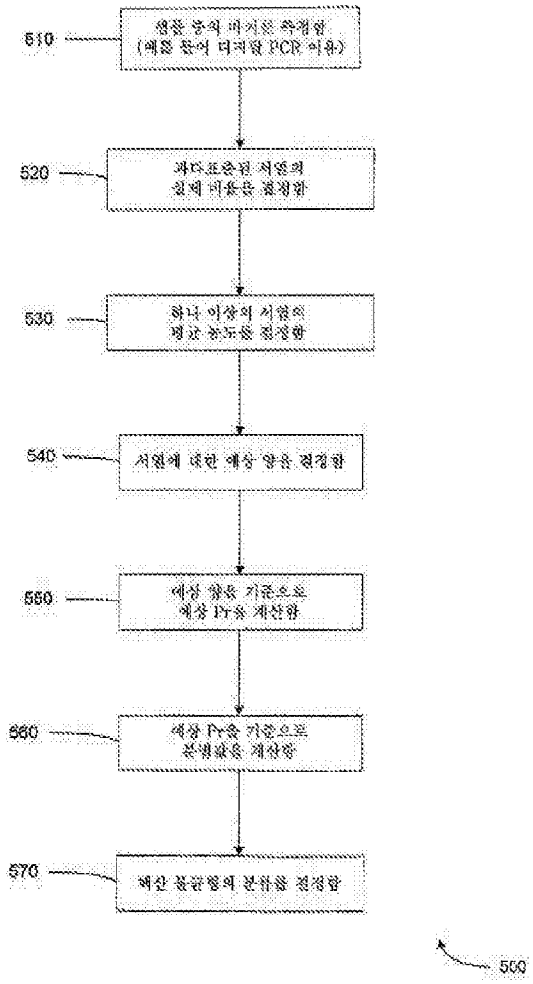
도면3


1.0
0.9
0.8
0.7
0.6
0.5
0.4
0
100
200
300
400
500
분류불가능
정해수성
정보제공 염의 수

도면4


410
샘플 중 태아 유전의 백분율을 측정함
420
디지털 PCR을 수행함
430
마커를 확인하고, 정보제공 염의 수를 결정함
440
과다표현형 서열의 실제 비율을 결정함
450
태아 유전의 백분율을 기준으로 문턱값을 계산함
460
Pr 값이 핵산 불균형을 나타내는지 여부를 결정함
400

도면5



도면6

m	정상 혈장 샘플에서 관찰된 변이의 평균 비율	파라프로운딩 서열과 관련된 변이의 평균 비율	두 분포 사이의 모든 비율에 대해 관찰된 평균 비율	정상 혈장 샘플에서 관찰된 변이의 평균 비율	파라프로운딩 서열과 관련된 변이의 평균 비율	로그값 (0.5 - 0.02) 계산	정규 시험 양의 계산	p_r
0.1	0.0955	0.1813	0.0173	0.0778	0.1554	3.13	0.0439	0.98
0.2	0.1813	0.3297	0.0598	0.1255	0.2699	3.22	0.1934	0.89
0.3	0.2552	0.4552	0.1169	0.1623	0.3545	3.31	0.4258	0.7
0.4	0.3207	0.5507	0.1815	0.1881	0.3959	3.46	0.6775	0.75
0.5	0.3975	0.6255	0.2487	0.1948	0.3834	3.65	0.8282	0.73
0.6	0.4512	0.6988	0.3133	0.1855	0.3533	3.82	0.9196	0.74
0.7	0.5034	0.7534	0.3703	0.1691	0.3193	3.93	0.9882	0.75
0.8	0.5507	0.7981	0.4198	0.1532	0.2886	3.98	0.9896	0.78
0.9	0.5939	0.8347	0.4651	0.0961	0.1794	3.46	0.8055	0.78
1	0.6321	0.8647	0.5066	0.0935	0.1733	3.75	0.8858	0.79
1.1	0.6671	0.8892	0.5403	0.0838	0.1556	3	0.7489	0.8
1.2	0.6988	0.9098	0.5694	0.0704	0.1318	4.72	0.9223	0.83
1.3	0.7275	0.9275	0.5934	0.054	0.1015	4.87	0.8663	0.82
1.4	0.7534	0.9395	0.6106	0.0459	0.2316	5.05	0.7774	0.83
1.5	0.7769	0.9503	0.6282	0.0387	0.2225	5.48	0.7307	0.85
1.6	0.7981	0.9595	0.6436	0.0315	0.1997	5.93	0.7265	0.86
1.7	0.8175	0.9656	0.59	0.0273	0.1736	5.47	0.7039	0.87
1.8	0.8347	0.9727	0.6115	0.0238	0.1688	7.02	0.8216	0.88
1.9	0.8504	0.9776	0.6514	0.019	0.1462	7.69	0.8672	0.88
2	0.8647	0.9817	0.6488	0.0151	0.1329	8.39	0.9457	0.89

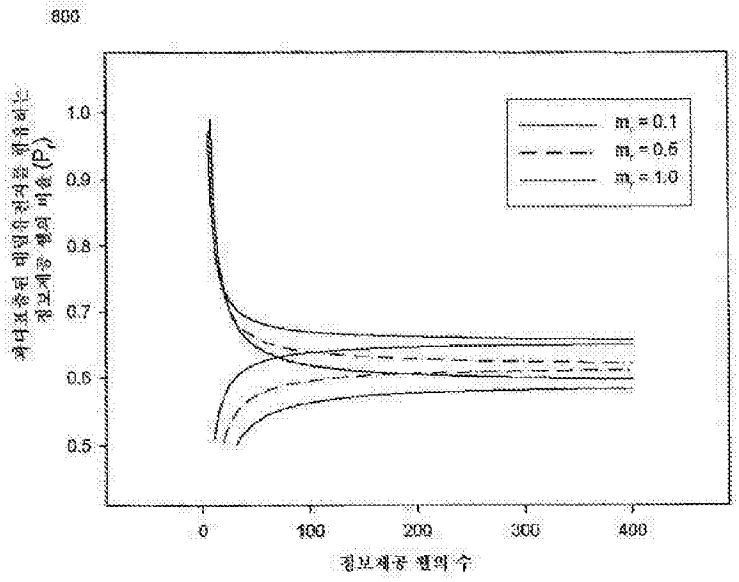
600

도면7

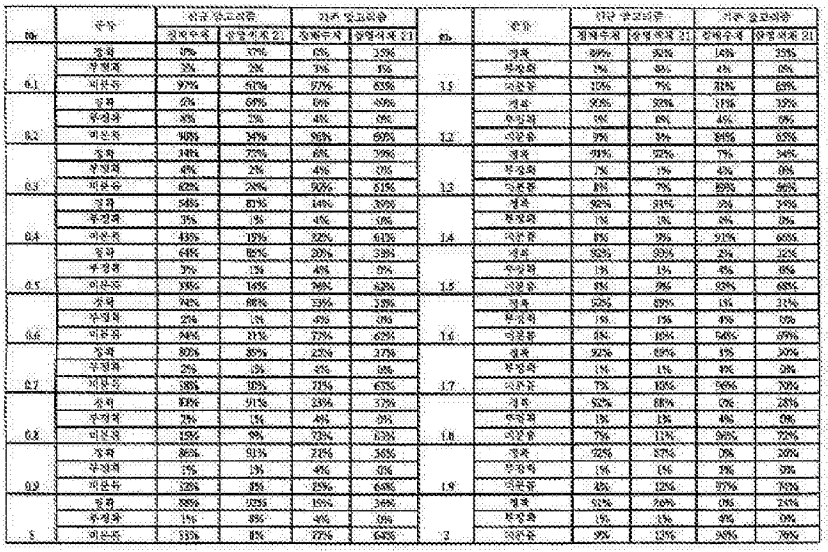
m_i	생존 중 태아 DNA의 백분율							
	10%		25%		50%		100%	
	정보제공 횟수 백분율	P_r	정보제공 횟수 백분율	P_r	정보제공 횟수 백분율	P_r	정보제공 횟수 백분율	P_r
0.1	18%	0.51	18%	0.53	19%	0.56	21%	0.61
0.2	30%	0.51	31%	0.53	32%	0.56	35%	0.61
0.3	39%	0.51	40%	0.53	41%	0.57	43%	0.62
0.4	45%	0.51	47%	0.54	48%	0.57	48%	0.63
0.5	48%	0.52	49%	0.54	49%	0.57	51%	0.63
0.6	50%	0.52	50%	0.54	50%	0.58	51%	0.64
0.7	50%	0.52	50%	0.54	50%	0.58	50%	0.65
0.8	49%	0.52	49%	0.54	49%	0.58	48%	0.65
0.9	48%	0.52	47%	0.55	47%	0.59	46%	0.66
1	46%	0.52	45%	0.55	44%	0.59	43%	0.67
1.1	44%	0.52	43%	0.55	42%	0.6	40%	0.68
1.2	41%	0.52	40%	0.55	39%	0.6	37%	0.69
1.3	39%	0.52	38%	0.55	36%	0.6	34%	0.69
1.4	36%	0.52	35%	0.56	33%	0.61	31%	0.7
1.5	34%	0.52	33%	0.56	31%	0.61	28%	0.71
1.6	31%	0.52	30%	0.56	28%	0.62	26%	0.72
1.7	29%	0.53	28%	0.56	26%	0.62	23%	0.73
1.8	27%	0.53	25%	0.57	24%	0.63	21%	0.73
1.9	24%	0.53	23%	0.57	21%	0.63	19%	0.74
2	22%	0.53	21%	0.57	20%	0.64	17%	0.75

700

도면8

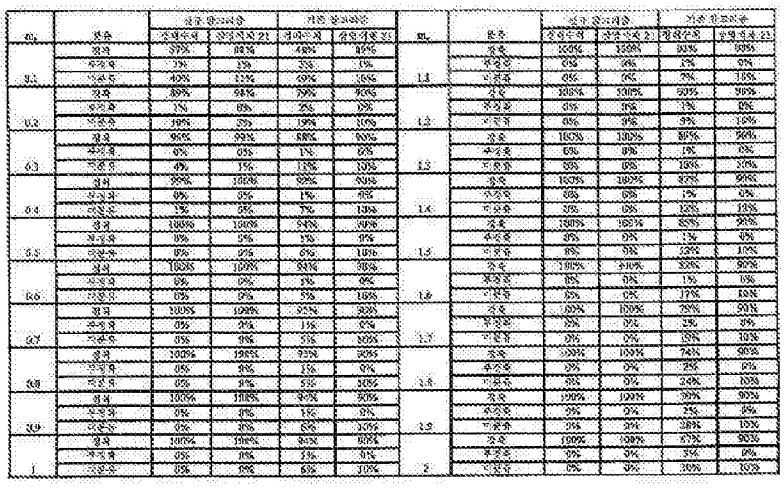


도면9a



900

도면9b



950

도면10

m_r	분류 결과	정해수치	T21	m_r	분류 결과	정해수치	T21
0.1	정확	57%	86%	0.9	정확	100%	100%
	부정확	3%	1%		부정확	0%	0%
	미분류	40%	11%		미분류	0%	0%
0.2	정확	89%	98%	1	정확	100%	100%
	부정확	1%	0%		부정확	0%	0%
	미분류	10%	2%		미분류	0%	0%
0.3	정확	96%	99%	1.1	정확	100%	100%
	부정확	0%	0%		부정확	0%	0%
	미분류	4%	1%		미분류	0%	0%
0.4	정확	99%	100%	1.2	정확	100%	100%
	부정확	0%	0%		부정확	0%	0%
	미분류	1%	0%		미분류	0%	0%
0.5	정확	100%	100%	1.3	정확	100%	100%
	부정확	0%	0%		부정확	0%	0%
	미분류	0%	0%		미분류	0%	0%
0.6	정확	100%	100%	1.4	정확	100%	100%
	부정확	0%	0%		부정확	0%	0%
	미분류	0%	0%		미분류	0%	0%
0.7	정확	100%	100%	1.5	정확	100%	100%
	부정확	0%	0%		부정확	0%	0%
	미분류	0%	0%		미분류	0%	0%
0.8	정확	100%	100%	2	정확	100%	100%
	부정확	0%	0%		부정확	0%	0%
	미분류	0%	0%		미분류	0%	0%

1000

도면11

m_r	분류 결과	정해수치	상원적재 21	m_r	분류 결과	정해수치	상원적재 21
0.1	정확	38%	44%	0.7	정확	95%	96%
	부정확	2%	2%		부정확	0%	1%
	미분류	60%	54%		미분류	5%	5%
0.2	정확	64%	69%	0.8	정확	95%	95%
	부정확	2%	2%		부정확	1%	0%
	미분류	34%	29%		미분류	4%	5%
0.3	정확	78%	82%	0.9	정확	96%	96%
	부정확	2%	1%		부정확	0%	0%
	미분류	20%	17%		미분류	4%	4%
0.4	정확	85%	87%	1	정확	97%	96%
	부정확	1%	1%		부정확	0%	0%
	미분류	14%	12%		미분류	3%	4%
0.5	정확	89%	90%	1.5	정확	97%	96%
	부정확	1%	1%		부정확	0%	0%
	미분류	10%	9%		미분류	3%	4%
0.6	정확	92%	93%	2	정확	96%	95%
	부정확	1%	1%		부정확	0%	0%
	미분류	7%	7%		미분류	4%	5%

1100

도면12

염색 수	분석 방식	샘플 중 염색 DNA 비율율							
		10%		25%		50%		100%	
		정확수치	정확수치 비율	정확수치	정확수치 비율	정확수치	정확수치 비율	정확수치	정확수치 비율
384	정확	0%	0%	8%	7%	47%	45%	89%	80%
	부정확	0%	0%	0%	1%	2%	2%	1%	1%
	미판독	100%	100%	93%	92%	51%	48%	10%	9%
768	정확	0%	0%	23%	26%	73%	76%	98%	99%
	부정확	0%	0%	2%	2%	2%	1%	0%	0%
	미판독	100%	100%	75%	74%	25%	23%	2%	1%
1152	정확	1%	1%	39%	40%	88%	87%	100%	100%
	부정확	0%	0%	2%	2%	1%	1%	0%	0%
	미판독	99%	99%	59%	58%	13%	12%	0%	0%
1536	정확	2%	2%	50%	51%	93%	93%	100%	100%
	부정확	0%	0%	2%	2%	1%	1%	0%	0%
	미판독	98%	97%	48%	47%	7%	6%	0%	0%
1920	정확	4%	4%	60%	60%	96%	96%	100%	100%
	부정확	0%	1%	2%	2%	0%	0%	0%	0%
	미판독	96%	95%	38%	37%	4%	3%	0%	0%
3840	정확	18%	19%	84%	83%	100%	100%	100%	100%
	부정확	1%	1%	1%	1%	0%	0%	0%	0%
	미판독	81%	80%	15%	15%	18%	22%	0%	0%
7680	정확	43%	45%	97%	97%	100%	100%	100%	100%
	부정확	2%	2%	0%	0%	0%	0%	0%	0%
	미판독	55%	53%	3%	3%	0%	0%	0%	0%

1200

도면13a

샘플	프로세싱 방*	염색 수	서열 위치유전자에 대해 염색된 염색 수				m_p	SPRT 결과		
			A 염색	G 염색	AG	모두 염색		P_0	동일성 가능한 염색	분류
정상 DNA										
N677	AG	384	88	83	125	93	0.79	0.51	0.63 - 0.86	정확수치
N710	AG	384	102	83	73	126	0.82	0.55	0.61 - 0.83	정확수치
N435	AGG	384	48	157	139	48	0.63	0.76	0.62 - 0.84	121
N681	AAG	384	135	69	82	98	0.5	0.66	0.61 - 0.83	121
암수 DNA										
V639	AG	384	109	89	71	117	0.80	0.53	0.61 - 0.83	정확수치
V645	AG	384	89	100	74	121	0.85	0.63	0.61 - 0.83	정확수치
N435	AGG	384	52	138	95	99	0.48	0.73	0.61 - 0.83	121
3215	AAG	384	140	56	136	52	0.71	0.72	0.62 - 0.84	121

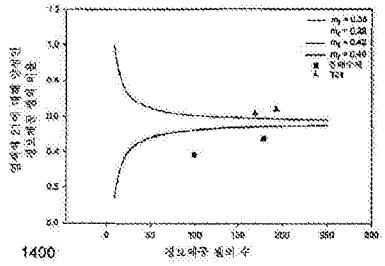
1300

도면13b

샘플	프로세싱	염색 수	서열 위치유전자에 대해 염색된 염색 수				m_p	SPRT 결과		
			A 염색	G 염색	AG	모두 염색		이상	동일성 가능한 염색	분류
M2360P	AG	384	90	100	97	97	0.67	0.528	0.61 - 0.84	정확수치
M2361P	AG	384	97	105	85	117	0.65	0.52	0.61 - 0.83	정확수치
M2473P	AG	384	96	92	94	102	0.8	0.682	0.62 - 0.82	정확수치
M2524P	AG	384	29	28	3	324	0.08	0.508	0.54 - 0.84	정확수치
M2528P	AG	384	112	85	44	143	0.41	0.668	0.60 - 0.82	정확수치
M2531P	AG	384	90	101	72	121	0.66	0.538	0.61 - 0.83	정확수치
M2532P	AG	384	73	81	57	183	0.41	0.536	0.60 - 0.83	정확수치
M2636P	AG	384	68	80	53	178	0.37	0.577	0.59 - 0.82	정확수치
M2639P	AG	384	73	86	17	240	0.21	0.689	0.58 - 0.82	정확수치
M2525P	AAG	384	110	83	21	200	0.21	0.675	0.59 - 0.81	121
M2722P	AAG	384	248	127	112	296	0.37	0.66	0.60 - 0.83	121
M2718P	AGG	384	95	114	99	139	0.46	0.638	0.60 - 0.82	121
M1519P	AGG	384	88	130	54	142	0.34	0.691	0.59 - 0.82	121

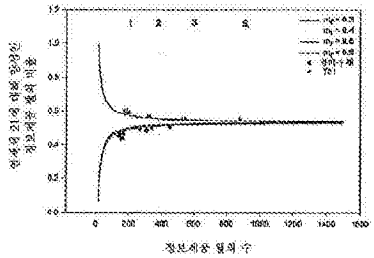
1350

도면14a



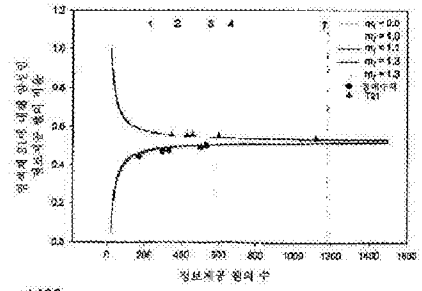
1400

도면14b



1430

도면14c



1460

도면15a

샘플	유전자형	횟수	개별 대립유전자에 대해 양성인 횟수				SPRT 결과			
			A 단독	G 단독	GA	모두 음성	m[a]	P[b]	불확정구간의 하한 - 상한	분류[c]
V243	AG	765	168	209	117	271	0.47	0.554	0.509- 0.621	정상수체
N435	AGG	765	92	281	105	287	0.30	0.753	0.597- 0.819	T21
T21-5	AAG	765	267	126	141	231	0.43	0.675	0.607- 0.758	T21

[a] 평균 염색 분자의 수

[b] P는 다음 식에 의해 계산하였음:

$$\frac{\text{자녀염색체 그룹을 선택한 경우 양성인 횟수}}{\text{A에 대해서만 양성인 횟수} + \text{G에 대해서만 양성인 횟수}}$$

[c] P<이 분류기능을 위해 미리 설정한 경우에 "정상수체"로 지정하였고; P>이 분류기능을 위해 설정한 경우에 "T21"로 지정하였음. T21, 삼염색체 21

도면15b

피험자	횟수 수	개별 대립유전자에 대해 양성인 횟수			
		A 단독	G 단독	GA	모두 음성
S1	765	14	15	0	736
S2	765	7	10	0	748
S3	765	13	9	1	744
S4	765	7	8	0	750
S5	765	12	8	0	745
S6	765	8	14	0	743
S7	765	13	9	0	743
S8	765	9	8	0	748
S9	765	14	15	0	736
S10	765	9	12	0	744
S11	765	8	17	0	740
S12	765	11	9	0	745

도면15c

샘플	유전자형	패널	개체 대립유전자에 대해 양성인 좌위 수					SPRT 분석		
			합계 수	A 단독	G 단독	GA	조기 중성	P_1	분류유용가능한 영역	분류
M2870P	AG	5	3825	82	103	5	3635	0.557	0.570- 0.602	정배수체
M2875P	AG	3	2295	35	40	0	2220	0.533	0.548- 0.625	정배수체
M3078P	AG	3	2295	33	33	1	2228	0.500	0.541- 0.631	정배수체
M2978P	AG	12	9180	56	46	0	9078	0.548	0.555- 0.618	정배수체
M2881P	AAG	2	1530	42	13	0	1475	0.754	0.631- 0.640	T21

* 분류에 필요한 패널의 수

도면16a

		A 또는 G 대립유전자에 대해 양성인 좌위 수				SPRT 분석		
샘플	유전자형*	A 단독	G 단독	둘 다	m^b	P_1^c	분류유용가능한 영역[d]	분류*
N3305	AG	157	151	48	0.27	0.545	0.59-0.61	정배수체
N3326	AG	174	132	52	0.28	0.559	0.59-0.61	정배수체
N3351	AG	146	125	28	0.23	0.531	0.58-0.60	정배수체
N3355	AG	149	135	46	0.27	0.525	0.59-0.61	정배수체
N2772	AG	139	149	79	0.35	0.516	0.60-0.61	정배수체
N3308	AG	154	145	66	0.32	0.513	0.60-0.61	정배수체
N3355	AG	94	73	7	0.11	0.563	0.58-0.61	정배수체
N3369	AG	101	109	28	0.18	0.519	0.59-0.61	정배수체
N3361	AG	91	78	17	0.12	0.568	0.58-0.61	정배수체
N828	AAG	153	89	22	0.16	0.632	0.58-0.66	T18
CA2	AGG	12	28	0	0.016	0.700	0.51-0.56	T18
CA7	AGG	97	171	36	0.23	0.638	0.59-0.61	T18

* 유전자형은 질량 분광법에 의해 결정하였음.

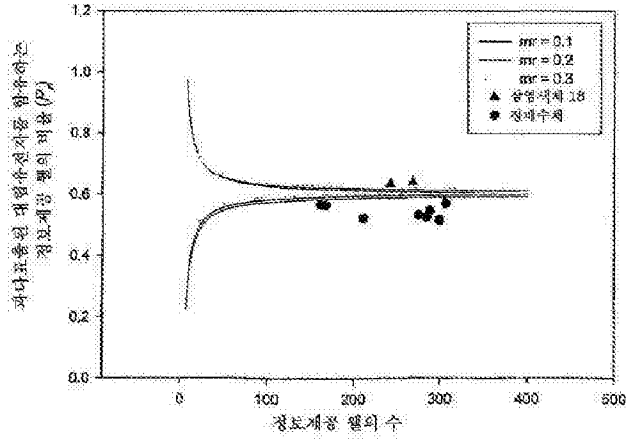
b m 값은 별 장 참조 부형 분석에 방금 수용 나타냄.

c P_1 값은 자식표출된 대립유전자에 대해 양성인 정도 지표 영역 비율을 나타냄.

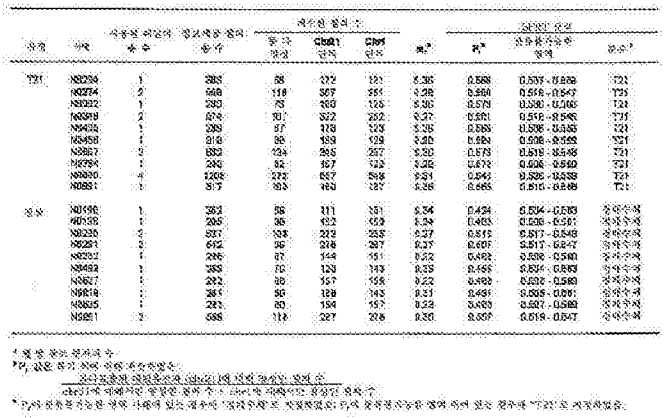
d 분류유용가능한 영역은 m 값에 따라 변함.

e P_1이 분류유용가능한 영역 내에 있는 경우에 '정배수체'로 지정하였고, P_1이 분류유용가능한 영역 외에 있는 경우에 'T18'로 지정하였음.

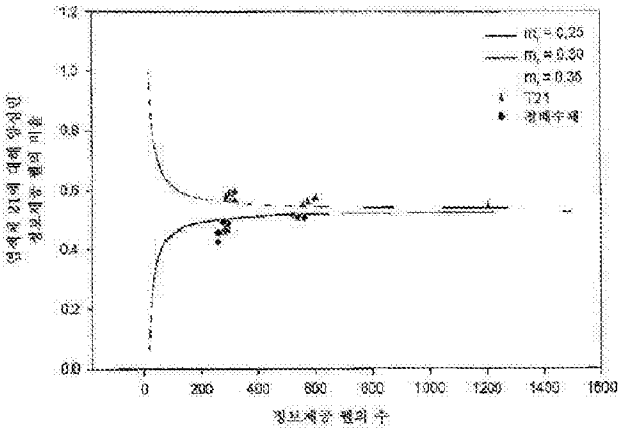
도면16b



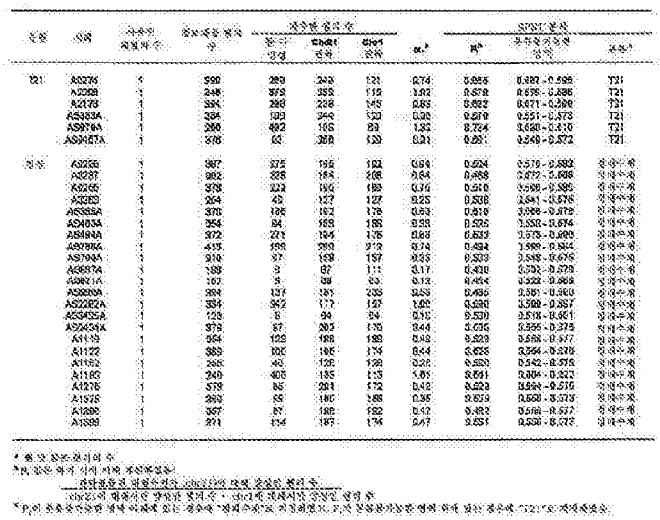
도면17



도면18



도면19



도면20

유형	샘플	사용된 체크포인트 수	검사수행 횟수	염색체 분석 수			α*	SPRT 분석		
				총 수	Chr18	Chr21		R_0*	판정중지구간	결과**
이수성	960230	1	360	242	151	199	0.85	0.475	0.573 - 0.589	E
	960635	1	370	225	176	191	0.78	0.464	0.570 - 0.586	E
	960851	1	358	206	192	166	0.67	0.538	0.564 - 0.583	E
T18	960629	1	543	236	309	214	0.86	0.605	0.578 - 0.590	T18
	960630	1	586	253	240	105	0.92	0.594	0.574 - 0.592	T18
	960636	1	572	385	339	183	0.88	0.583	0.573 - 0.590	E
	961876	1	349	115	221	122	0.37	0.544	0.561 - 0.574	T18
	961701	1	363	54	191	112	0.31	0.536	0.546 - 0.575	T18

* 총 검출로 종료된 수

** 판정 중지 시점 염색 체수 계산식(일):

$$\frac{\text{기준염색체 내염색체수 (chr18)에 검출된 염색 수}}{\text{chr21에 염색체인 검출된 염색 수 + chr18에 염색체인 검출된 염색 수}}$$

*** 참조 분류불가능한 경우 이외의 값은 추후에 "염색체수"로 치환하였으며, 참조 분류불가능한 경우 역시 같은 경우에 "T18"로 치환하였음.

도면21

유형	샘플	염색체의 수	검정	검사수행 횟수	염색체 분석 수			α*	SPRT 분석		
					총 수	Chr 21	염색 염색		R_0*	판정중지구간	결과**
정상	960356	1	chr21/chr1	1	79	77	77	0.53	0.480	0.486 - 0.580	E
		1	chr21/chr18	1	82	79	78	0.54	0.520	0.487 - 0.584	N
			총합		161	156	155	0.46	0.500	0.512 - 0.560	E
	960358	1	chr21/chr1	1	73	70	62	0.56	0.432	0.482 - 0.581	E
		1	chr21/chr18	1	66	60	60	0.50	0.508	0.488 - 0.575	N
			총합		141	130	122	0.55	0.472	0.507 - 0.557	E
	960359	1	chr21/chr1	1	59	51	51	0.45	0.500	0.480 - 0.581	N
		1	chr21/chr18	1	45	41	40	0.41	0.481	0.488 - 0.582	N
			총합		104	92	91	0.43	0.495	0.512 - 0.559	E
	960361	1	chr21/chr1	1	63	61	52	0.47	0.468	0.486 - 0.578	E
		1	chr21/chr18	1	61	60	54	0.45	0.532	0.480 - 0.583	N
			총합		124	121	106	0.43	0.490	0.514 - 0.560	E
	960362	1	chr21/chr1	1	97	101	85	0.49	0.501	0.508 - 0.575	N
		1	chr21/chr18	1	81	81	68	0.77	0.512	0.505 - 0.567	N
			총합		178	182	153	0.48	0.505	0.522 - 0.557	N
		2	chr21/chr1	1	124	120	98	0.68	0.488	0.503 - 0.577	E
			총합		302	301	251	0.46	0.513	0.505 - 0.552	E
21	960357	1	chr21/chr1	1	100	98	100	0.51	0.500	0.500 - 0.573	N
		1	chr21/chr18	1	104	98	90	0.50	0.517	0.495 - 0.579	N
			총합		204	196	190	0.50	0.508	0.510 - 0.556	N
		2	chr21/chr1	1	78	75	68	0.53	0.497	0.495 - 0.576	N
			총합		282	271	257	0.51	0.502	0.502 - 0.545	N
		2	chr21/chr18	1	92	88	83	0.57	0.504	0.495 - 0.575	N
			총합		296	283	240	0.53	0.502	0.507 - 0.545	N
		3	chr21/chr1	1	164	158	141	0.56	0.509	0.504 - 0.575	N
			총합		411	396	430	0.54	0.508	0.505 - 0.545	N
		3	chr21/chr18	1	102	100	98	0.53	0.508	0.502 - 0.573	T21
			총합		1599	478	515	0.55	0.500	0.501 - 0.544	T21
	960332	1	chr21/chr1	1	104	109	79	0.58	0.500	0.500 - 0.577	T21
		1	chr21/chr18	1	101	98	81	0.55	0.520	0.501 - 0.574	T21
			총합		204	191	213	0.56	0.500	0.510 - 0.557	T21
	960365	1	chr21/chr1	1	102	93	76	0.49	0.525	0.496 - 0.582	N
		1	chr21/chr18	1	99	87	78	0.53	0.504	0.499 - 0.586	N
			총합		221	180	154	0.43	0.525	0.512 - 0.568	N
		2	chr21/chr1	1	103	98	85	0.41	0.507	0.495 - 0.585	N
			총합		423	374	268	0.51	0.504	0.513 - 0.563	N
		2	chr21/chr18	1	108	43	100	0.54	0.505	0.496 - 0.568	T21
			총합		661	217	355	0.49	0.500	0.503 - 0.566	T21
	960794	1	chr21/chr1	1	102	92	105	53	0.502	0.496 - 0.575	T21
		1	chr21/chr18	1	105	85	97	51	0.548	0.498 - 0.575	N
			총합		208	166	202	104	0.504	0.515 - 0.559	T21
	960595	1	chr21/chr1	1	100	105	108	72	0.50	0.500	0.500 - 0.576	T21
		1	chr21/chr18	1	108	86	104	64	0.50	0.500	0.499 - 0.574	N
			총합		208	171	212	136	0.50	0.500	0.515 - 0.558	T21

* 총 검출로 종료된 수

** 판정 중지 시점 염색 체수 계산식(일):

$$\frac{\text{기준염색체 내염색체수 (chr18)에 검출된 염색 수}}{\text{chr21에 염색체인 검출된 염색 수 + chr18에 염색체인 검출된 염색 수}}$$

*** 참조 분류불가능한 경우 이외의 값은 추후에 "염색체수"로 치환하였으며, 참조 분류불가능한 경우 역시 같은 경우에 "T21"로 치환하였음.

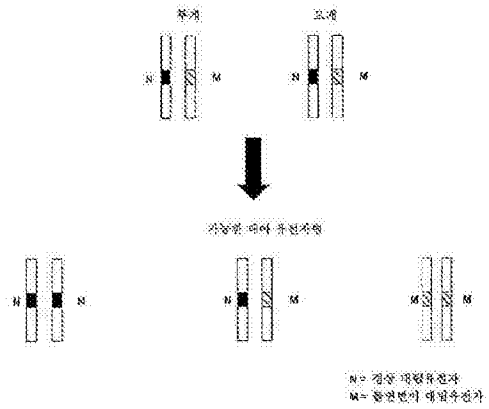
도면22a

		조이상 프랑킨 홀 수		카	분류
		영역에 21	정보 영역에		
정해주제 사례 1	관찰 1	37	33	0.523468751	이분류 (Decision)
	관찰 2	31	42	0.435063276	이분류
	관찰 3	29	37	0.44198795	이분류
	관찰 4	39	26	0.6008691	이분류
	중해이프1 소계	135	138	0.485621	이분류
	관찰 1	32	29	0.56365316	이분류
	관찰 2	27	24	0.53165424	이분류
	관찰 3	10	17	0.3681535	이분류
	관찰 4	24	16	0.6529754	이분류
	중해이프 2 소계	93	84	0.52474	이분류
	총계	239	223	0.507182	정해주제
사례 2	관찰 1	46	35	0.563215685	이분류
	관찰 2	31	45	0.41246525	이분류
	관찰 3	22	22	0.59554763	이분류
	관찰 4	35	35	0.5328175	이분류
	중해이프1 소계	146	135	0.519153	이분류
	관찰 1	28	23	0.55272135	이분류
	관찰 2	23	23	0.51196833	이분류
	관찰 3	25	15	0.66858157	이분류
	관찰 4	29	26	0.5293349	이분류
	중해이프 2 소계	102	89	0.545204	이분류
	중해이프 1+2 소계	248	221	0.529583	이분류
	관찰 1	17	25	0.4122146	이분류
	관찰 2	25	16	0.6027289	이분류
	관찰 3	13	26	0.3382602	이분류
	관찰 4	27	24	0.5316872	이분류
	중해이프 3 소계	82	91	0.475775	이분류
	중해이프 1~3 소계	331	311	0.515369	이분류
	관찰 1	15	9	0.6259193	이분류
	관찰 2	14	15	0.4824155	이분류
	관찰 3	18	17	0.5146313	이분류
	관찰 4	21	19	0.5279080	이분류
	중해이프 4 소계	69	60	0.522354	이분류
	중해이프 1~4 소계	399	372	0.517942	이분류

도면22b

	관찰 1	10	13	0.4337957	이분류
	관찰 2	11	14	0.4360082	이분류
	관찰 3	7	19	0.3018022	이분류
	관찰 4	19	10	0.6579085	이분류
	중해이프 5 소계	46	54	0.469951	이분류
	총계	447	426	0.5133444	정해주제
T21					
사례 3	관찰 1	50	33	0.60012269	이분류
	관찰 2	56	47	0.54458053	이분류
	관찰 3	46	28	0.6136525	이분류
	관찰 4	59	36	0.6106731	이분류
	총계	205	144	0.59079	T21
사례 4	관찰 1	56	36	0.61065513	이분류
	관찰 2	38	32	0.53605912	이분류
	관찰 3	40	24	0.62788068	이분류
	관찰 4	47	27	0.634511	이분류
	총계	180	119	0.603475	T21

도면23



도면24a

사례	생물	예상된 %	재발의 수	ZFY	ZFX	ALL POS	NET ZFY	NET ZFX	참조 대립유전자	m_r	SPRT[a]
1	XY + XX	50%	1	239	465	159	64	310	Y	0.37	XX
	(XX %)	25%	1	328	447	205	123	242	Y	0.56	XX
2	$XY_1 + XY_2$	50%	1	414	405	220	194	185	X	0.85	XY
	(XY_2 %)	25%	1	432	428	245	187	181	X	0.81	XY

도면24b

사례	재발의 수	ZFY	ZFX	ALL POS	NET ZFY	NET ZFX	참조 대립유전자	m_r	SPRT
1	1	81	171	18	63	153	Y	0.11	XX
2	1	103	175	23	80	152	Y	0.14	XX
3	1	110	175	28	82	147	Y	0.16	XX
4	1	111	158	24	87	132	Y	0.18	XX
5	2	236	322	50	186	272	Y	0.17	XX
6	1	116	164	29	87	155	Y	0.18	XX
7	1	94	169	19	75	150	Y	0.13	XX
8	1	91	155	24	67	131	Y	0.13	XX
9	3	333	447	74	259	373	Y	0.18	XX
10	1	110	154	24	86	130	Y	0.15	XX
11	1	83	170	26	57	144	Y	0.11	XX
12	1	85	159	18	69	143	Y	0.12	XX

표 25

사례	체크의 수	모계 유전자형[1]	부계 유전자형[2]	대립유전자 개수 합	A	B	ALL PDB	NST A	NST B	신호 강도	유의성	유의성	PL	SPRT[3]
1	1	AB	BB	1355	208	516	116	148	302	30	A		0.42	BB
2	1	AB	BB	1355	208	379	81	127	185	14.0	A		0.32	BB
3	1	AB	AB	1355	249	251	37	161	183	16.4	A		0.59	AB
4	1	AB	BB	1355	158	348	91	157	188	18.7	A		0.23	BB
5	1	AB	AB	1355	329	249	73	156	176	16.5	A		0.55	AB
6	1	AB	AB	1355	318	249	139	156	188	15.1	A		0.54	AB
7	1	AB	AB	1355	354	278	84	164	195	16.8	A		0.49	AB
8	1	AB	BB	1355	296	363	144	152	214	15.7	A		0.39	BB
9	1	AB	AB	1355	179	385	109	179	185	13.6	A		0.45	AB
10	1	AB	AA	1355	397	273	120	217	180	NN	G		0.66	AA
1	1	AB	BB	2355	248	286	85	152	198	13.0	A		0.28	BB
2	1	AB	BB	2355	199	286	79	118	219	18.5	A		0.53	BB
3	1	AB	AB	2355	248	218	72	172	143	16.0	G		0.52	H
4	1	AB	BB	2355	158	287	99	99	195	25.5	A		0.23	BB
5	1	AB	AB	2355	331	215	71	169	144	25.0	G		0.53	AB
6	1	AB	AB	2355	281	280	98	183	159	28.5	A		0.46	AB
7	1	AB	AB	2355	358	231	88	186	143	25.7	G		0.26	AB
8	1	AB	BB	2355	248	311	104	142	207	23.6	A		0.35	BB
9	1	AB	AB	2355	267	249	83	184	166	21.9	G		0.39	AB
10	1	AB	AA	2355	398	202	77	238	126	NN	G		0.61	AA
1	1	AB	BB	2355	344	291	89	146	205	15.7	A		0.38	BB
2	1	AB	BB	2355	288	314	88	131	226	24.0	A		0.31	BB
3	1	AB	AB	2355	354	213	55	149	157	18.6	A		0.51	AB
4	1	AB	BB	2355	193	385	78	103	211	28.9	A		0.29	BB
5	1	AB	AB	2355	333	243	77	152	166	24.5	A		0.56	AB
6	1	AB	AB	2355	292	263	110	182	153	27.0	A		0.48	AB
7	1	AB	AB	2355	28	275	95	186	159	22.5	A		0.43	BB
8	1	AB	BB	2355	226	327	56	148	251	28.7	A		0.27	BB
9	1	AB	AB	2355	371	260	104	197	178	25.2	A		0.44	AB
10	1	AB	AA	2355	343	184	48	167	106	NN	G		0.22	AA
1	1	AB	BB	5055	109	386	97	79	288	33.9	A		0.20	BB
2	1	AB	BB	5055	163	334	76	93	285	45.8	A		0.21	BB
3	1	AB	AB	5055	280	209	55	147	155	65.7	A		0.36	AB
4	1	AB	BB	5055	193	386	80	73	300	58.9	A		0.19	BB
5	1	AB	AB	5055	241	248	51	156	155	57.7	G		0.39	AB
6	1	AB	AB	5055	214	285	68	100	177	53.6	A		0.34	AB
7	1	AB	BB	5055	265	283	103	152	173	51.3	A		0.41	AB
8	1	AB	BB	5055	158	336	82	73	218	57.6	A		0.23	BB
9	1	AB	AB	5055	335	229	74	189	151	55.7	G		0.35	AB
10	1	AB	AA	5055	317	158	55	202	71	NN	G		0.18	AA

[1] A = 모계염색체의 대립유전자, G = 이중염색체형성.
[2] 부계 체크 DNA 또는 각각의 ZFYX 증폭산물의 크기에 따라 결정.
[3] 대립유전자(s), 각각의 ZFYX 증폭은 모두 사용가능함.
[4] SPRT는 시퀀싱을 계속해서 각각의 ZFYX 염기서열 개수가 모두 정해진 개수 또는 수렴하였을 때, 더 이상.

표 26a

사례	모계 유전자형[1]	부계 유전자형[2]	체크의 수	M	N	ALL PDB	NST M	NST N	1 신호 강도 염기 개수	1 SPRT	2 대립형 ZFYX간의 염기 개수	2 SPRT	동조 대립유전자	PL
1	MN	MM	5	649	388	272	573	469	5%	MM	0.5%	MM	N	0.48
	MN	MM	2	853	748	288	574	458	90%	MM	0.5%	MM	N	0.39
	MN	MM	1	323	259	78	244	140	25%	MM	24.5%	MM	N	0.34
	MN	MM	1	418	180	82	328	89	80%	MM	57.5%	MM	N	0.23
2	MN	MM	9	2869	2008	881	1872	1589	5%	M	3.5%	M	N	0.45
	MN	MM	3	835	607	382	535	567	10%	MM	9.7%	MM	M	0.49
	MN	MM	1	277	255	67	186	168	25%	MM	24.5%	MM	M	0.42
	MN	MM	1	273	295	101	172	194	30%	NN	58.0%	MM	M	0.46

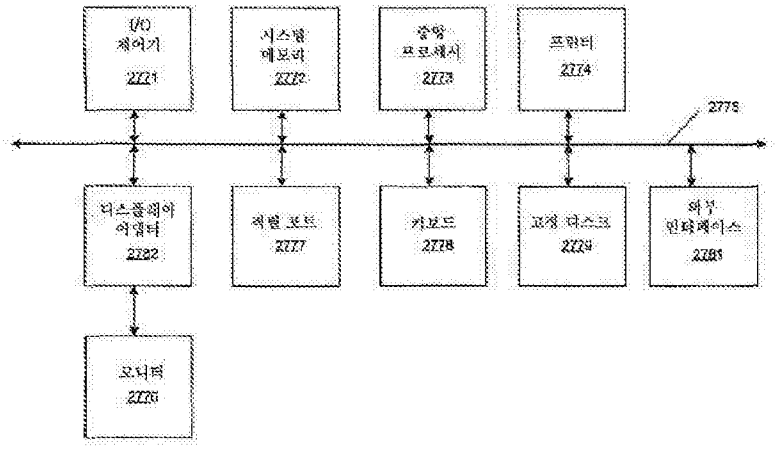
[1] M = CTTT 증폭 존재, N = CTTT 증폭 부재.
[2] SPRT는 각각을 별도 수행하였음.
[3] SPRT는 각각의 ZFYX 증폭으로 얻어진 염기 개수 수렴하였을 때, 더 이상.

표 26b

사례	체크의 수	모계 유전자형[1]	부계 유전자형[2]	대립유전자 개수 합	M	N	ALL PDB	NST M	NST N	신호 강도 염기 개수	유의성	PL	SPRT[3]
1	1	MN	MM	25%	203	206	66	146	142	29.0	M	0.30	MM
2	1	MN	MN	25%	149	204	39	110	165	29.7	M	0.32	MM
3	1	MN	NN	25%	128	212	31	97	181	32.7	M	0.16	NN
4	1	MN	MM	25%	173	196	45	128	151	27.0	M	0.26	MM
5	1	MN	NN	25%	147	230	47	100	183	24.3	M	0.21	NN
6	1	MN	MN	25%	139	206	45	104	161	19.2	M	0.30	MN
7	1	MN	NN	25%	185	237	58	107	179	18.7	M	0.24	NN
8	1	MN	MN	25%	203	180	43	160	137	23.4	N	0.27	MN
9	1	MN	NN	25%	173	241	61	112	182	16.6	M	0.26	NN
10	1	MN	MN	25%	174	190	52	122	138	16.8	M	0.28	U
11	1	MN	NN	25%	197	242	48	118	183	17.7	M	0.15	NN
12	1	MN	MN	25%	213	189	47	166	152	21.8	N	0.30	MN

[1] M = CTTT 증폭 존재, N = CTTT 증폭 부재.
[2] 신호 강도 DNA 또는 각각의 ZFYX 산정으로 결정함.
[3] SPRT는 각각의 ZFYX 산정으로 얻어진 염기 개수 수렴하였을 때, U, 더 이상.

도면27



서 열 목 록

SEQUENCE LISTING

<110> Lo, Yuk-Ming Dennis

Chiu, Rossa Wai Kwun

Chan, Kwan Chee

Zee, Benny Chung Ying

Chong, Ka Chun

The Chinese University of Hong Kong

<120> Determining a Nucleic Acid Sequence mbalance

<130> IPA161076-CN

<140> WO PCT/GB08/02524

<141> 2008-07-23

<150> US 60/951,438

<151> 2007-07-23

<160> 58

<170> FastSEQ for Windows Version 4.0

<210> 1

<211> 27

<212> DNA

<213> Artificial Sequence

<220>

<223> synthetic real-time PCR and digital RNA-SNP analysis

PLAC4 gene-specific reverse transcription primer

<400> 1

agtatataga accatgttta ggccaga 27

<210> 2

<211> 21

<212> DNA

<213> Artificial Sequence

<220>

<223> synthetic real-time PCR PLAC4 primer

<400> 2

ccgctagggt gtcttttaag c 21

<210> 3

<211> 25

<212> DNA

<213> Artificial Sequence

<220>

<223> synthetic real-time PCR PLAC4 primer

<400> 3

gtgttgcaat acaaaatgag tttct 25

<210> 4

<211> 14

<212> DNA

<213> Artificial Sequence

<220>

<223> synthetic real-time PCR PLAC4 fluorescent probe

<220>

<221> modified_base

<222> (1)...(1)

<223> a modified by 6-carboxyfluorescein (FAM)

<220>

<221> modified_base

<222> (14)...(14)

<223> c modified by minor groove binding non-fluorescent
quencher (MGBNFQ)

<400> 4

attggagcaa attc 14

<210> 5

<211> 104

<212> DNA

<213> Artificial Sequence

<220>

<223> single-stranded synthetic DNA oligonucleotide
amplicon

<400> 5

cgccgctagg gtgtctttta agctattgga gcaaattcaa atttggctta aagaaaaaga 60

aactcatttt gtattgcaac accaggagta tcccaaggga ctcg 104

<210> 6

<211> 22

<212> DNA

<213> Artificial Sequence

<220>

<223> synthetic digital RNA-SNP analysis non-intron
spanning forward primer

<400> 6

tttgtattgc aacaccattt gg 22

<210> 7

<211> 15

<212> DNA

<213> Artificial Sequence

<220>

<223> synthetic allele-G-specific TaqMan probe targeting
rs8130833 SNP of PLAC4

<220>

<221> modified_base

<222> (1)...(1)

<223> t modified by 6-carboxyfluorescein (FAM)

<220>

<221> modificd_basc

<222> (15)...(15)

<223> g modified by minor groove binding non-fluorescent quencher (MGBNFQ)

<400> 7

tcgtcgtcta acttg 15

<210> 8

<211> 17

<212> DNA

<213> Artificial Sequence

<220>

<223> synthetic allele-A-specific TaqMan probe targeting rs8130833 SNP of PLAC4

<220>

<221> modified_base

<222> (1)...(1)

<223> a modified by fluorescent reporter VIC

<220>

<221> modified_base

<222> (17)...(17)

<223> g modified by minor groove binding non-fluorescent quencher (MGBNFQ)

<400> 8

attcgtcatc taacttg 17

<210> 9

<211> 23

<212> DNA

<213> Artificial Sequence

<220>

<223> synthetic digital RCD analysis co-amplification forward primer for paralogous loci on chromosome 21 and 1

<400> 9

gttgttctgc aaaaaacctt cga 23

<210> 10

<211> 28

<212> DNA

<213> Artificial Sequence

<220>

<223> synthetic digital RCD analysis co-amplification
reverse primer for paralogous loci on chromosome
21 and 1

<400> 10

cttggccaga aatacttcat taccatat 28

<210> 11

<211> 19

<212> DNA

<213> Artificial Sequence

<220>

<223> synthetic chromosome-specific TaqMan probe
designed to target chromosome 21 paralog

<220>

<221> modified_base

<222> (1)...(1)

<223> t modified by 6-carboxyfluorescein (FAM)

<220>

<221> modified_base

<222> (19)...(19)

<223> a modified by minor groove binding non-fluorescent
quencher (MGBNFQ)

<400> 11

tacctccata atgagtaaa 19

<210> 12

<211> 20

<212> DNA

<213> Artificial Sequence

<220>

<223> synthetic chromosome-specific TaqMan probe
designed to target chromosome 1 paralog
<220>
<221> modified_base
<222> (1)...(1)
<223> c modified by fluorescent reporter VIC
<220>
<221> modified_base
<222> (20)...(20)
<223> a modified by minor groove binding non-fluorescent
quencher (MGBNFQ)
<400> 12
cgtacctctg taatgtgtaa 20
<210> 13
<211> 27
<212> DNA
<213> Artificial Sequence
<220>
<223> synthetic digital PCR PLAC4 gene-specific reverse

transcription primer
<400> 13
agtatataga accatgttta ggccaga 27
<210> 14
<211> 22
<212> DNA
<213> Artificial Sequence
<220>
<223> synthetic digital PCR amplification forward primer
for rs8130833 SNP region of PLAC4 gene
<400> 14
tttgtattgc aacaccattt gg 22
<210> 15
<211> 15
<212> DNA

<213> Artificial Sequence

<220>

<223> synthetic allele-G-specific TaqMan probe targeting
PLAC4 rs8130833 SNP

<220>

<221> modified_base

<222> (1)...(1)

<223> t modified by 6-carboxyfluorescein (FAM)

<220>

<221> modified_base

<222> (15)...(15)

<223> g modified by minor groove binding non-fluorescent
quencher (MGBNFQ)

<400> 15

tcgtcgtcta acttg 15

<210> 16

<211> 17

<212> DNA

<213> Artificial Sequence

<220>

<223> synthetic allele-A-specific TaqMan probe targeting
PLAC4 rs8130833 SNP

<220>

<221> modified_base

<222
> (1)...(1)

<223> a modified by fluorescent reporter VIC

<220>

<221> modified_base

<222> (17)...(17)

<223> g modified by minor groove binding non-fluorescent
quencher (MGBNFQ)

<400> 16

attcgtcatc taacttg 17

<210> 17

<211> 20

<212> DNA

<213> Artificial Sequence

<220>

<223> synthetic digital PCR amplification serpin
peptidase inhibitor clade B (ovalbumin) member 2
(SERPINB2) gene-specific reverse transcription
primer

<400> 17

cgcagacttc tcaccaaaca 20

<210> 18

<211> 20

<212> DNA

<213> Artificial Sequence

<220>

<223> synthetic digital PCR amplification serpin
peptidase inhibitor clade B (ovalbumin) member 2
(SERPINB2) forward primer

<400> 18

ctcagctctg caatcaatgc 20

<210> 19

<211> 16

<212> DNA

<213> Artificial Sequence

<220>

<223> synthetic allele-A-specific TaqMan probe targeting
serpin peptidase inhibitor clade B (ovalbumin)
member 2 (SERPINB2) rs6098 SNP

<220>

<221> modified_base

<222> (1)...(1)

<223> c modificd by 6-carboxyfluorcsccin (FAM)

<220>

<221> modified_base

<222> (16)...(16)

<223> t modified by minor groove binding non-fluorescent quencher (MGBNFQ)

<400> 19

ccacagggaa ttattt 16

<210> 20

<211> 16

<212> DNA

<213> Artificial Sequence

<220>

<223> synthetic allele-G-specific TaqMan probe targeting serpin peptidase inhibitor clade B (ovalbumin) member 2 (SERPINB2) rs6098 SNP

<220>

<221> modified_base

<222> (1)...(1)

<223> c modified by 6-carboxyfluorescein (FAM)

<220>

<221> modified_base

<222> (16)...(16)

<223> t modified by minor groove binding non-fluorescent quencher (MGBNFQ)

<400> 20

ccacagggga ttattt 16

<210> 21

<211> 23

<212> DNA

<213> Artificial Sequence

<220>

<223> synthetic digital PCR analysis co-amplification forward primer for 101-bp amplicon for paralogous

loci on chromosome 21 and 1

<400> 21

gttgttctgc aaaaaacctt cga 23

<210> 22

<211> 28

<212> DNA

<213> Artificial Sequence

<220>

<223> synthetic digital PCR analysis co-amplification reverse primer for 101-bp amplicon for paralogous loci on chromosome 21 and 1

<400> 22

cttggccaga aatacttcat taccatat 28

<210> 23

<211> 19

<212> DNA

<213> Artificial Sequence

<220>

<223> synthetic chromosome-specific TaqMan probe designed to target chromosome 21 paralog

<220>

<221> modified_base

<222> (1)...(1)

<223> t modified by 6-carboxyfluorescein (FAM)

<220>

<221> modified_base

<222> (19)...(19)

<223> a modified by minor groove binding non-fluorescent quencher (MGBNFQ)

<400> 23

tacctccata atgagtaaa 19

<210> 24

<211> 20

<212> DNA

<213> Artificial Sequence

<220>

<223> synthetic chromosome-specific TaqMan probe
designed to target chromosome 1 paralog

<220>

<221> modified_base

<222> (1)...(1)

<223> c modified by fluorescent reporter VIC

<220>

<221> modified_base

<222> (20)...(20)

<223> a modified by minor groove binding non-fluorescent
quencher (MGBNFQ)

<400> 24

cgtacctctg taatgtgtaa 20

<210> 25

<211> 24

<212> DNA

<213> Artificial Sequence

<220>

<223> synthetic digital RCD analysis co-amplification

forward primer for 128-bp amplicon for paralogous
loci on chromosome 21 and 18

<400> 25

gtacagaaac cacaaactga tcgg 24

<210> 26

<211> 17

<212> DNA

<213> Artificial Sequence

<220>

<223> synthetic digital RCD analysis co-amplification
reverse primer for 128-bp amplicon for paralogous
loci on chromosome 21 and 18

<400> 26

gtccaggctg tgggcct 17

<210> 27

<211> 14

<212> DNA

<213> Artificial Sequence

<220>

<223> synthetic chromosome-specific TaqMan probe
designed to target chromosome 21 paralog

<220>

<221> modified_base

<222> (1)...(1)

<223> a modified by 6-carboxyfluorescein (FAM)

<220>

<221> modified_base

<222> (14)...(14)

<223> a modified by minor groove binding non-fluorescent
quencher (MGBNFQ)

<400> 27

aagaggcgag gcaa 14

<210> 28

<211> 15

<212> DNA

<213> Artificial Sequence

<220>

<223> synthetic chromosome-specific TaqMan probe
designed to target chromosome 18 paralog

<220>

<221> modified_base

<222> (1)...(1)

<223> a modified by fluorescent reporter VIC

<220>

<221> modificd_basc

<222> (15)...(15)

<223> c modified by minor groove binding non-fluorescent quencher (MGBNFQ)

<400> 28

aagaggacag gcaac 15

<210> 29

<211> 33

<212> DNA

<213> Artificial Sequence

<220>

<223> synthetic digital PCR analysis co-amplification forward primer for 121-bp amplicon for paralogous loci on chromosome 21 and 1

<400> 29

acgttggatg gttgttctgc aaaaaacctt cga 33

<210> 30

<211> 38

<212> DNA

<213> Artificial Sequence

<220>

<223> synthetic digital PCR analysis co-amplification reverse primer for 121-bp amplicon for paralogous loci on chromosome 21 and 1

<400> 30

acgttggatg cttggccaga aatacttcat taccatat 38

<210> 31

<211> 21

<212> DNA

<213> Artificial Sequence

<220>

<223> synthetic digital PCR analysis co-amplification extension primer for 121-bp amplicon for paralogous loci on chromosome 21 and 1

<400> 31

ctcatcctca cttcgtacct c 21

<210> 32

<211> 34

<212> DNA

<213> Artificial Sequence

<220>

<223> synthetic digital RCD analysis co-amplification forward primer for 148-bp amplicon for paralogous loci on chromosome 21 and 18

<400> 32

acgttggatg gtacagaaac cacaaactga tcgg 34

<210> 33

<211> 27

<212> DNA

<213> Artificial Sequence

<220>

<223> synthetic digital RCD analysis co-amplification reverse primer for 148-bp amplicon for paralogous loci on chromosome 21 and 18

<400> 33

acgttggatg gtccaggctg tgggcct 27

<210> 34

<211> 17

<212> DNA

<213> Artificial Sequence

<220>

<223> synthetic digital RCD analysis co-amplification extension primer for 148-bp amplicon for paralogous loci on chromosome 21 and 18

<400> 34

acaaaagggg gaagagg 17

<210> 35

<211> 35

<212> DNA

<213> Artificial Sequence

<220>

<223> synthetic digital RCD analysis co-amplification
forward primer for 81 bp amplicon for paralogous
loci on chromosome 21 and 1

<400> 35

acgttggatg ttgatgaagt ctcatctcta cttcg 35

<210> 36

<211> 33

<212> DNA

<213> Artificial Sequence

<220>

<223> synthetic digital RCD analysis co-amplification
reverse primer for 81 bp amplicon for paralogous

loci on chromosome 21 and 1

<400> 36

acgttggatg caataagctt ggccagaaat act 33

<210> 37

<211> 37

<212> DNA

<213> Artificial Sequence

<220>

<223> synthetic digital RCD analysis co-amplification
forward primer for 82 bp amplicon for paralogous
loci on chromosome 21 and 7

<400> 37

acgttggatg gaatttaagc taaatcagcc tgaactg 37

<210> 38

<211> 37

<212> DNA

<213> Artificial Sequence

<220>

<223> synthetic digital RCD analysis co-amplification

reverse primer for 82 bp amplicon for paralogous loci on chromosome 21 and 7

<400> 38

acgttggatg gtttctcata gttcatcgta ggcttat 37

<210> 39

<211> 30

<212> DNA

<213> Artificial Sequence

<220>

<223> synthetic digital RCD analysis co-amplification

forward primer for 101 bp amplicon for paralogous loci on chromosome 21 and 2

<400> 39

acgttggatg tcaggcaggg ttctatgcag 30

<210> 40

<211> 30

<212> DNA

<

213> Artificial Sequence

<220>

<223> synthetic digital RCD analysis co-amplification

reverse primer for 101 bp amplicon for paralogous loci on chromosome 21 and 2

<400> 40

acgttggatg aggcggcttc ctggctcttg 30

<210> 41

<211> 31

<212> DNA

<213> Artificial Sequence

<220>

<223> synthetic digital RCD analysis co-amplification

forward primer for 102 bp amplicon for paralogous loci on chromosome 21 and 6

<400> 41

acgttggatg gctcgtctca ggctcgtagt t 31

<210> 42

<211> 30

<212> DNA

<213> Artificial Sequence

<220>

<223> synthetic digital RCD analysis co-amplification reverse primer for 102 bp amplicon for paralogous loci on chromosome 21 and 6

<400> 42

acgttggatg tttcttcgag cccttcttgg 30

<210> 43

<211> 23

<212> DNA

<213> Artificial Sequence

<220>

<223> synthetic extension primer for paralogous sequence mismatches (PSM) on paralogous loci on chromosomes 21 and 1

<400

> 43

gtctcatctc tacttcgtac ctc 23

<210> 44

<211> 19

<212> DNA

<213> Artificial Sequence

<220>

<223> synthetic extension primer for paralogous sequence mismatches (PSM) on paralogous loci on chromosomes 21 and 7

<400> 44

ttttacgctg tccccattt 19

<210> 45

<211> 19

<212> DNA

<213> Artificial Sequence

<220>

<223> synthetic extension primer for paralogous sequence mismatches (PSM) on paralogous loci on chromosomes 21 and 2

<400> 45

ggtctatgca ggagccgac 19

<210> 46

<211> 25

<212> DNA

<213> Artificial Sequence

<220>

<223> synthetic extension primer for paralogous sequence mismatches (PSM) on paralogous loci on chromosomes 21 and 6

<400> 46

tgggcgcggg agcggacttc gctgg 25

<210> 47

<211> 25

<212> DNA

<213> Artificial Sequence

<220>

<223> synthetic digital PCR analysis co-amplification forward primer for 87-bp amplicon for zinc finger protein sequences on chromosome X (ZFX) and Y (ZFY)

<400> 47

caagtgctgg actcagatgt aactg 25

<210> 48

<211> 28

<212> DNA

<213> Artificial Sequence

<220>

<223> synthetic digital PCR analysis co-amplification reverse primer for 87-bp amplicon for zinc finger protein sequences on chromosome X (ZFX) and Y (ZFY)

<400> 48

tgaagtaatg tcagaagcta aaacatca 28

<210> 49

<211> 16

<212> DNA

<213> Artificial Sequence

<220>

<223> synthetic chromosome-specific TaqMan probe designed to target chromosome X paralog

<220>

<221> modified_base

<222> (1)...(1)

<223> t modified by fluorescent reporter VIC

<220>

<221> modified_base

<222> (16)...(16)

<223> a modified by minor groove binding non-fluorescent quencher (MGBNFQ)

<400> 49

tctttagcac attgca 16

<210> 50

<211> 17

<212

> DNA

<213> Artificial Sequence

<220>

<223> synthetic chromosome-specific TaqMan probe designed to target chromosome Y paralog

<220>

<221> modified_base

<222> (1)...(1)

<223> t modified by 6-carboxyfluorescein (FAM)

<220>

<221> modified_base

<222> (17)...(17)

<223> c modified by minor groove binding non-fluorescent
quencher (MGBNFQ)

<400> 50

tctttaccac actgcac 17

<210> 51

<211> 19

<212> DNA

<213> Artificial Sequence

<220>

<223> synthetic digital PCR co-amplification forward
primer for 87-bp amplicon of beta-globin
(hemoglobin beta (HBB)) HbE mutation forward
primer

<400> 51

gggcaaggtg aacgtggat 19

<210> 52

<211> 29

<212> DNA

<213> Artificial Sequence

<220>

<223> synthetic digital PCR co-amplification forward
primer for 87-bp amplicon of beta-globin
(hemoglobin beta (HBB)) HbE mutation reverse
primer

<400> 52

ctattggtct ccttaaacct gtcttgtaa 29

<210> 53

<211> 14

<212> DNA

<213> Artificial Sequence

<220>

<223> synthetic allele-G-specific TaqMan probe targeting beta-globin (hemoglobin beta (HBB)) HbE mutation normal allele

<220>

<221> modified_base

<222> (1)...(1)

<223> t modified by fluorescent reporter VIC

<220>

<221> modified_base

<222> (14)...(14)

<223> c modified by minor groove binding non-fluorescent quencher (MGBNFQ)

<400> 53

ttggtggtga ggcc 14

<210> 54

<211> 14

<212> DNA

<213> Artificial Sequence

<220>

<223> synthetic allele-A-specific TaqMan probe targeting beta-globin (hemoglobin beta (HBB)) HbE mutation mutant allele

<220>

<221> modified_base

<222> (1)...(1)

<223> t modified by 6-carboxyfluorescein (FAM)

<220>

<221> modified_base

<222> (14)...(14)

<223> c modified by minor groove binding non-fluorescent quencher (MGBNFQ)

<400> 54

ttggtggtaa ggcc 14

<210> 55

<211> 20

<212> DNA

<213> Artificial Sequence

<220>

<223> synthetic digital PCR co-amplification forward primer for 87-bp and 83-bp amplicon of CD41/42 deletion mutation normal and mutant allele

<400> 55

ttttcccacc cttaggctgc 20

<210> 56

<211> 23

<212> DNA

<213> Artificial Sequence

<220>

<223> synthetic digital PCR co-amplification reverse primer for 87-bp and 83-bp amplicon of CD41/42 deletion mutation normal and mutant allele

<400> 56

acagcatcag gagtggacag atc 23

<210> 57

<211> 19

<212> DNA

<213> Artificial Sequence

<220>

<223> synthetic allele-specific TaqMan probe targeting CD41/42 deletion mutation normal allele (without dclction)

<220>

<221> modified_base

<222> (1)...(1)

<223> c modified by fluorescent reporter VIC

<220>

<221> modified_base

<222> (19)...(19)

<223> t modified by minor groove binding non-fluorescent quencher (MGBNFQ)

<400> 57

cagaggttct ttgagtcct 19

<210> 58

<211> 15

<212> DNA

<213> Artificial Sequence

<220>

<223> synthetic allele-specific TaqMan probe targeting CD41/42 deletion mutation mutant allele (with deletion)

<220>

<221> modified_base

<222> (1)...(1)

<223> a modified by 6-carboxyfluorescein (FAM)

<220>

<221> modified_base

<222> (15)...(15)

<223> t modified by minor groove binding non-fluorescent quencher (MGBNFQ)

<400> 58

agaggttgag tcctt 15

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

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
29 January 2009 (29.01.2009)

PCT

(10) International Publication Number
WO 2009/013492 A1

(51) International Patent Classification:
G06F 19/00 (2006.01) *C12Q 1/68* (2006.01)

(21) International Application Number:
PCT/GB2008/002524

(22) International Filing Date: 23 July 2008 (23.07.2008)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
60/951,438 23 July 2007 (23.07.2007) US

(71) Applicant *(for all designated States except US)*: **THE CHINESE UNIVERSITY OF HONG KONG** [CN/CN]; Technology Licensing Office, Room 226, Pi Ch'iu Building, Shatin, New Territories (HK).

(72) Inventors; and
(75) Inventors/Applicants *(for US only)*: **LO, Yuk-Ming Dennis** [GB/CN]; 4th Floor, 7 King Tak Street, Homantin, Kowloon (HK). **CHIU, Rossa Wai Kwun** [AU/CN]; Flat 1A, Block 1, Constellation Cove, 1 Hung Lam Drive, Tai Po, New Territories (HK). **CHAN, Kwan Chee** [CN/CN]; Flat A, 13/F, Block 34, Broadway Street, Mei Foo Sun Chuen, Kowloon (HK). **ZEE, Benny Chung Ying** [CA/CN]; Flat 18E, Tower 2, La Costa, Ma on Shan, New Territories (HK). **CHONG, Ka Chun** [CN/CN]; Flat 06, 29/F, Shin King House, Fu Shin Estate, Tai Po, New Territories (HK).

(74) Agents: **BUFTON, Karen** et al.; Mewburn Ellis LLP, 33 Gutter Lane, London EC2V 8AS (GB).

(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, 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, PG, PH, PL, PT,

[Continued on next page]

(54) Title: DETERMINING A NUCLEIC ACID SEQUENCE IMBALANCE

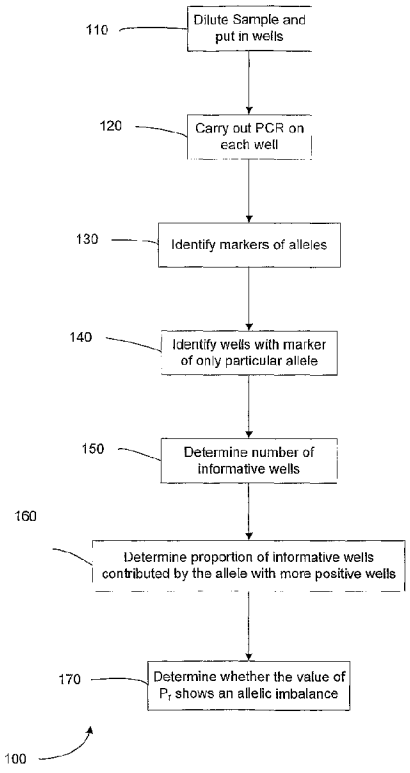


FIG. 1

(57) Abstract: Methods, systems, and apparatus are provided for determining whether a nucleic acid sequence imbalance exists within a biological sample. One or more cutoff values for determining an imbalance of, for example, the ratio of the two sequences (or sets of sequences) are chosen. The cutoff value may be determined based at least in part on the percentage of fetal DNA in a sample, such as maternal plasma, containing a background of maternal nucleic acid sequences. The cutoff value may also be determined based on an average concentration of a sequence per reaction. In one aspect, the cutoff value is determined from a proportion of informative wells that are estimated to contain a particular nucleic acid sequence, where the proportion is determined based on the above-mentioned percentage and/or average concentration. The cutoff value may be determined using many different types of methods, such as sequential probability ratio testing (SPRT).

WO 2009/013492 A1

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, MT, NL, NO, PL, PT, RO, SE, SI, SK, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Published:
— *with international search report*

DETERMINING A NUCLEIC ACID SEQUENCE IMBALANCE

CLAIM OF PRIORITY

[0001] The present application claims priority from and is a non-provisional application of U.S. Provisional Application No. 60/951438, entitled "DETERMINING A NUCLEIC ACID SEQUENCE IMBALANCE" filed July 23, 2007 (Attorney Docket No. 016285-005200US), the entire contents of which are herein incorporated by reference for all purposes.

CROSS-REFERENCES TO RELATED APPLICATIONS

[0002] The present application is also related to concurrently filed non-provisional application entitled "DIAGNOSING FETAL CHROMOSOMAL ANEUPLOIDY USING GENOMIC SEQUENCING," (Attorney Docket No. 016285-005220US), the entire contents of which are herein incorporated by reference for all purposes.

FIELD OF THE INVENTION

[0003] This invention generally relates to the diagnostic testing of genotypes and diseases by determining an imbalance between two different nucleic acid sequences, and more particularly to the identification of Down syndrome, other chromosomal aneuploidies, mutations and genotypes in a fetus via testing a sample of maternal blood. The invention also relates to the detection of cancer, the monitoring of transplantation, and the monitoring of infectious diseases.

BACKGROUND

[0004] Genetic diseases, cancers, and other conditions often result from or produce an imbalance in two corresponding chromosomes or alleles or other nucleic acid sequences. That is an amount of one sequence relative to another sequence is larger or smaller than normal. Usually, the normal ratio is an even 50/50 ratio. Down Syndrome (trisomy 21) is such a disease having an imbalance of an extra chromosome 21.

[0005] Conventional prenatal diagnostic methods of trisomy 21 involve the sampling of fetal materials by invasive procedures such as amniocentesis or chorionic villus sampling, which pose a finite risk of fetal loss. Non-invasive procedures, such as screening by ultrasonography and biochemical markers, have been used to risk-stratify pregnant women prior to definitive invasive diagnostic procedures. However, these screening methods typically measure epiphenomena that are associated with trisomy 21 instead of the core

chromosomal abnormality, and thus have suboptimal diagnostic accuracy and other disadvantages, such as being highly influenced by gestational age.

[0006] The discovery of circulating cell-free fetal DNA in maternal plasma in 1997 offered new possibilities for noninvasive prenatal diagnosis (Lo, YMD and Chiu, RWK 2007 *Nat Rev Genet* 8, 71-77). While this method has been readily applied to the prenatal diagnosis of sex-linked (Costa, JM et al. 2002 *N Engl J Med* 346, 1502) and certain single gene disorders (Lo, YMD et al. 1998 *N Engl J Med* 339, 1734-1738), its application to the prenatal detection of fetal chromosomal aneuploidies has represented a considerable challenge (Lo, YMD and Chiu, RWK 2007, supra). First, fetal nucleic acids co-exist in maternal plasma with a high background of nucleic acids of maternal origin that can often interfere with the analysis (Lo, YMD et al. 1998 *Am J Hum Genet* 62, 768-775). Second, fetal nucleic acids circulate in maternal plasma predominantly in a cell-free form, making it difficult to derive dosage information of genes or chromosomes within the fetal genome.

[0007] Significant developments overcoming these challenges have recently been made (Benachi, A & Costa, JM 2007 *Lancet* 369, 440-442). One approach detects fetal-specific nucleic acids in the maternal plasma, thus overcoming the problem of maternal background interference (Lo, YMD and Chiu, RWK 2007, supra). Dosage of chromosome 21 was inferred from the ratios of polymorphic alleles in the placenta-derived DNA/RNA molecules. However, this method is less accurate when samples contain lower amount of the targeted gene and can only be applied to fetuses who are heterozygous for the targeted polymorphisms, which is only a subset of the population if one polymorphism is used.

[0008] Dhallan et al (Dhallan, R, *et al.* 2007, supra Dhallan, R, *et al.* 2007 *Lancet* 369, 474-481) described an alternative strategy of enriching the proportion of circulating fetal DNA by adding formaldehyde to maternal plasma. The proportion of chromosome 21 sequences contributed by the fetus in maternal plasma was determined by assessing the ratio of paternally-inherited fetal-specific alleles to non-fetal-specific alleles for single nucleotide polymorphisms (SNPs) on chromosome 21. SNP ratios were similarly computed for a reference chromosome. An imbalance of fetal chromosome 21 was then inferred by detecting a statistically significant difference between the SNP ratios for chromosome 21 and those of the reference chromosome, where significant is defined using a fixed p-value of ≤ 0.05. To ensure high population coverage, more than 500 SNPs were targeted per chromosome. However, there have been controversies regarding the effectiveness of formaldehyde to

enrich to a high proportion (Chung, GTY, et al. 2005 *Clin Chem* 51, 655-658), and thus the reproducibility of the method needs to be further evaluated. Also, as each fetus and mother would be informative for a different number of SNPs for each chromosome, the power of the statistical test for SNP ratio comparison would be variable from case to case (Lo, YMD & Chiu, RWK. 2007 *Lancet* 369, 1997). Furthermore, since these approaches depend on the detection of genetic polymorphisms, they are limited to fetuses heterozygous for these polymorphisms.

[0009] Using polymerase chain reaction (PCR) and DNA quantification of a chromosome 21 locus and a reference locus in amniocyte cultures obtained from trisomy 21 and euploid fetuses, Zimmermann et al (2002 *Clin Chem* 48, 362-363) were able to distinguish the two groups of fetuses based on the 1.5-fold increase in chromosome 21 DNA sequences in the former. Since a 2-fold difference in DNA template concentration constitutes a difference of only one threshold cycle (Ct), the discrimination of a 1.5-fold difference has been the limit of conventional real-time PCR. To achieve finer degrees of quantitative discrimination, alternative strategies are needed. Accordingly, some embodiments of the present invention use digital PCR (Vogelstein, B et al. 1999 *Proc Natl Acad Sci U S A* 96, 9236-9241) for this purpose.

[0010] Digital PCR has been developed for the detection of allelic ratio skewing in nucleic acid samples (Chang, HW *et al.* 2002 *J Natl Cancer Inst* 94, 1697-1703). Clinically, it has been shown to be useful for the detection of loss of heterozygosity (LOH) in tumor DNA samples (Zhou, W. *et al.* 2002 *Lancet* 359, 219-225). For the analysis of digital PCR results, sequential probability ratio testing (SPRT) has been adopted by previous studies to classify the experimental results as being suggestive of the presence of LOH in a sample or not (El Karoui at al. 2006 *Stat Med* 25, 3124-3133). In methods used in the previous studies, the cutoff value to determine LOH used a fixed reference ratio of the two alleles in the DNA of 2/3. As the amount, proportion, and concentration of fetal nucleic acids in maternal plasma are variable, these methods are not suitable for detecting trisomy 21 using fetal nucleic acids in a background of maternal nucleic acids in maternal plasma.

[0011] It is desirable to have a noninvasive test for fetal trisomy 21 (and other imbalances) detection based on circulating fetal nucleic acid analysis, especially one that is independent of the use of genetic polymorphisms and/or of fetal-specific markers. It is also desirable to have accurate determination of cutoff values and counting of sequences, which can reduce the

number of wells of data and/or the amount of maternal plasma nucleic acid molecules necessary for accuracy, thus providing increased efficiency and cost-effectiveness. It is also desirable that noninvasive tests have high sensitivity and specificity to minimize false diagnoses.

[0012] Another application for fetal DNA detection in maternal plasma is for the prenatal diagnosis of single gene disorders such as beta-thalassemia. However, as fetal DNA only constitutes a minor fraction of DNA in maternal plasma, it is thought that this approach can only detect a mutation that a fetus has inherited from its father, but which is absent from the mother. Examples of this include the 4 bp deletion in codon 41/42 of the beta-globin gene causing beta-thalassemia (Chiu RWK et al. 2002 *Lancet*, 360, 998-1000) and the Q890X mutation of the cystic fibrosis transmembrance conductance regulator gene causing cystic fibrosis (Gonzalez-Gonzalez et al 2002 *Prenat Diagn,* 22, 946-8). However, as both beta-thalassemia and cystic fibrosis are autosomal recessive conditions, in which the fetus would need to inherit a mutation from each parent before the disease would manifest itself, the detection of merely the paternally-inherited mutation would only increase the risk of having the fetus having the disease from 25% to 50%. Diagnostically this is not ideal. Thus, the main diagnostic application of the existing approach would be for the scenario when no paternally-inherited fetal mutation can be detected in maternal plasma, when the fetus can then be excluded from having the homozygous disease state. However, diagnostically, this approach has the disadvantage that the conclusion is made based on the negative detection of the paternal mutation. Thus, an approach which would allow the complete fetal genotype (be it homozygous normal, homozygous mutant, or heterozygous) to be determined from maternal plasma, without the above limitation, would be very desirable.

BRIEF SUMMARY

[0013] Embodiments of this invention provides methods, systems, and apparatus for determining whether a nucleic acid sequence imbalance (e.g., allelic imbalance, mutational imbalance, or chromosome imbalance) exists within a biological sample. One or more cutoff values for determining an imbalance of, for example, a ratio of amounts of the two sequences (or sets of sequences) are chosen.

[0014] In one embodiment, the cutoff value is determined based at least in part on the percentage of fetal (clinically relevant nucleic acid) sequences in a biological sample, such as maternal plasma or serum or urine, which contains a background of maternal nucleic acid

sequences. In another embodiment, the cutoff value is determined based on an average concentration of a sequence in a plurality of reactions. In one aspect, the cutoff value is determined from a proportion of informative wells that are estimated to contain a particular nucleic acid sequence, where the proportion is determined based on the above-mentioned percentage and/or average concentration.

[0015] The cutoff value may be determined using many different types of methods, such as SPRT, false discovery, confidence interval, receiver operating characteristic (ROC). This strategy further minimized the amount of testing required before confident classification could be made. This is of particular relevance to plasma nucleic acid analysis where the template amount is often limiting.

[0016] According to one exemplary embodiment, a method is provided for determining whether a nucleic acid sequence imbalance exists within a biological sample, the method comprising: receiving data from a plurality of reactions, wherein the data includes: (1) a first set of quantitative data indicating a first amount of a clinically relevant nucleic acid sequence; and (2) a second set of quantitative data indicating a second amount of a background nucleic acid sequence different from the clinically relevant nucleic acid sequence; determining a parameter from the two data sets; deriving a first cutoff value from an average concentration of a reference nucleic acid sequence in each of the plurality of reactions, wherein the reference nucleic acid sequence is either the clinically relevant nucleic acid sequence or the background nucleic acid sequence; comparing the parameter to the first cutoff value; and based on the comparison, determining a classification of whether a nucleic acid sequence imbalance exists.

[0017] According to another exemplary embodiment, a method is provided for determining whether a nucleic acid sequence imbalance exists within a biological sample, the method comprising: receiving data from a plurality of reactions, wherein the data includes: (1) a first set of quantitative data indicating a first amount of a clinically relevant nucleic acid sequence; and (2) a second set of quantitative data indicating a second amount of a background nucleic acid sequence different from the clinically relevant nucleic acid sequence, wherein the clinically relevant nucleic acid sequence and the background nucleic acid sequence come from a first type of cells and from one or more second types of cells; determining a parameter from the two data sets; deriving a first cutoff value from a first percentage resulting from a measurement of an amount of a nucleic acid sequence from the first type of cells in the

biological sample; comparing the parameter to the cutoff value; and based on the comparison, determining a classification of whether a nucleic acid sequence imbalance exists.

[0018] Other embodiments of the invention are directed to systems and computer readable media associated with methods described herein.

[0019] A better understanding of the nature and advantages of the present invention may be gained with reference to the following detailed description and the accompanying drawings.

BRIEF DESCRIPTION OF THE DRAWINGS

[0020] FIG. 1 is a flowchart illustrating a digital PCR experiment.

[0021] FIG. 2A illustrates a digital RNA-SNP and RCD method according to an embodiment of the present invention.

[0022] FIG. 2B shows a table of examples of frequently detectable chromosomal aberrations in cancers.

[0023] FIG. 3 illustrates a graph having SPRT curves used to determine Down syndrome according to an embodiment of the present invention.

[0024] FIG. 4 shows a method of determining a disease state using a percentage of fetal cells according to an embodiment of the present invention.

[0025] FIG. 5 shows a method of determining a disease state using an average concentration according to an embodiment of the present invention.

[0026] FIG. 6 shows a table that tabulates the expected digital RNA-SNP allelic ratio and P_r of trisomy 21 samples for a range of template concentrations expressed as the average reference template concentration per well (m_r) according to an embodiment of the present invention.

[0027] FIG. 7 shows a table that tabulates the expected P_r for the fractional fetal DNA concentrations of 10%, 25%, 50% and 100% in trisomy 21 samples at a range of template concentrations expressed as the average reference template concentration per well (m_r) according to an embodiment of the present invention.

[0028] FIG. 8 shows a plot illustrating the degree of differences in the SPRT curves for m_r values of 0.1, 0.5 and 1.0 for digital RNA-SNP analysis according to an embodiment of the present invention.

[0029] FIG. 9A shows a table of a comparison of the effectiveness of the new and old SPRT algorithms for classifying euploid and trisomy 21 cases in 96-well digital RNA-SNP analyses according to an embodiment of the present invention.

[0030] FIG. 9B shows a table of a comparison of the effectiveness of the new and old SPRT algorithms for classifying euploid and trisomy 21 cases in 384-well digital RNA-SNP analyses according to an embodiment of the present invention.

[0031] FIG. 10 is a table showing the percentages of fetuses correctly and incorrectly classified as euploid or aneuploid and those not classifiable for the given informative counts according to an embodiment of the present invention.

[0032] FIG. 11 is a table 1100 showing computer simulations for digital RCD analysis for a pure (100%) fetal DNA sample according to an embodiment of the present invention.

[0033] FIG. 12 is a table 1200 showing results of computer simulation of accuracies of digital RCD analysis at m_f=0.5 for the classification of samples from euploid or trisomy 21 fetuses with different fractional concentrations of fetal DNA according to an embodiment of the present invention.

[0034] FIG. 13A shows a table 1300 of digital RNA-SNP analysis in placental tissues of euploid and trisomy 21 pregnancies according to an embodiment of the present invention.

[0035] FIG. 13B shows a table 1350 of digital RNA-SNP analysis of maternal plasma from euploid and trisomy 21 pregnancies according to an embodiment of the present invention.

[0036] FIG. 14A-14C show plots illustrating a cutoff curve resulting from an RCD analysis according to an embodiment of the present invention.

[0037] FIG. 15A shows a table of digital RNA-SNP analysis in placental tissues of euploid and trisomy 21 pregnancies according to an embodiment of the present invention.

[0038] FIG. 15B shows a table of digital RNA-SNP data of the 12 reaction panels from one maternal plasma sample according to an embodiment of the present invention.

[0039] FIG. 15C shows a table of digital RNA-SNP analysis of maternal plasma from euploid and trisomy 21 pregnancies according to an embodiment of the present invention.

[0040] FIG. 16A shows a table for a digital RNA-SNP analysis of euploid and trisomy 18 placentas according to an embodiment of the present invention.

[0041] FIG. 16B shows an SPRT interpretation of digital RNA-SNP data for euploid and trisomy 18 placentas according to an embodiment of the present invention.

[0042] FIG. 17 shows a table of a digital RCD analysis of 50% placental/maternal blood cell DNA mixtures of euploid and trisomy 21 pregnancies according to an embodiment of the present invention.

[0043] FIG. 18 shows a SPRT curve illustrating the decision boundaries for correct classification according to an embodiment of the present invention.

[0044] FIG. 19 shows a table of digital RCD analysis of amniotic fluid samples from euploid and trisomy 21 pregnancies according to an embodiment of the present invention.

[0045] FIG. 20 shows a table of digital RCD analysis of placental DNA samples from euploid and trisomy 18 pregnancies (E= euploid; T18= trisomy 18) according to an embodiment of the present invention.

[0046] FIG. 21 shows a table of a multiplex digital RCD analysis of 50% placental/maternal blood cell DNA mixtures of euploid and trisomy 21 pregnancies (E= euploid; T21= trisomy 21; U= unclassified) according to an embodiment of the present invention.

[0047] FIGS. 22A and 22B show a table of a multiplex digital RCD analysis of 50% euploid or trisomy 21 placental genomic DNA/50% maternal buffy coat DNA mix according to an embodiment of the present invention. Unclass denotes unclassifiable and T21 denotes trisomy 21.

[0048] FIG. 23 shows a scenario where both the male and female partners carry the same mutation.

[0049] FIG. 24A shows a table of a digital RMD analysis of female/male and male/male DNA mixtures according to an embodiment of the present invention.

[0050] FIG. 24B shows a table of a digital RMD analysis of mixtures with 25% female and 75% male DNA according to an embodiment of the present invention.

[0051] FIG. 25 shows a table of a digital RMD analysis of 15%-50% DNA mixtures mimicking maternal plasma samples for HbE mutation according to an embodiment of the present invention.

[0052] FIG. 26A shows a table of a digital RMD analysis of 5%-50% DNA mixtures mimicking maternal plasma samples for CD41/42 mutation according to an embodiment of the present invention.

[0053] FIG. 26B shows a table of a digital RMD analysis of 20% DNA mixtures mimicking maternal plasma samples for CD41/42 mutation according to an embodiment of the present invention.

[0054] FIG. 27 shows a block diagram of an exemplary computer apparatus usable with system and methods according to embodiments of the present invention.

DEFINITIONS

[0055] The term "*biological sample*" as used herein refers to any sample that is taken from a subject (*e.g.*, a human, such as a pregnant woman) and contains one or more nucleic acid molecule(s)of interest.

[0056] The term "*nucleic acid*" or "*polynucleotide*" refers to a deoxyribonucleic acid (DNA) or ribonucleic acid (RNA) and a polymer thereof in either single- or double-stranded form. Unless specifically limited, the term encompasses nucleic acids containing known analogs of natural nucleotides that have similar binding properties as the reference nucleic acid and are metabolized in a manner similar to naturally occurring nucleotides. Unless otherwise indicated, a particular nucleic acid sequence also implicitly encompasses conservatively modified variants thereof (*e.g.*, degenerate codon substitutions), alleles, orthologs, SNPs, and complementary sequences as well as the sequence explicitly indicated. Specifically, degenerate codon substitutions may be achieved by generating sequences in which the third position of one or more selected (or all) codons is substituted with mixed-base and/or deoxyinosine residues (Batzer *et al.*, *Nucleic Acid Res.* **19**:5081 (1991); Ohtsuka *et al.*, *J. Biol. Chem.* **260**:2605-2608 (1985); and Rossolini *et al.*, *Mol. Cell. Probes* 8:91-98 (1994)). The term nucleic acid is used interchangeably with gene, cDNA, mRNA, small noncoding RNA, micro RNA (miRNA), Piwi-interacting RNA, and short hairpin RNA (shRNA) encoded by a gene or locus.

[0057] The term "*gene*" means the segment of DNA involved in producing a polypeptide chain. It may include regions preceding and following the coding region (leader and trailer) as well as intervening sequences (introns) between individual coding segments (exons).

[0058] The term "*reaction*" as used herein refers to any process involving a chemical, enzymatic, or physical action that is indicative of the presence or absence of a particular polynucleotide sequence of interest. An example of a "*reaction*" is an amplification reaction such as a polymerase chain reaction (PCR). Another example of a "*reaction*" is a sequencing reaction, either by synthesis or by ligation. An "*informative reaction*" is one that indicates the presence of one or more particular polynucleotide sequence of interest, and in one case where only one sequence of interest is present. The term "*well*" as used herein refers to a reaction at a predetermined location within a confined structure, e.g., a well-shaped vial, cell, or chamber in a PCR array.

[0059] The term "*clinically relevant nucleic acid sequence*" as used herein can refer to a polynucleotide sequence corresponding to a segment of a larger genomic sequence whose potential imbalance is being tested or to the larger genomic sequence itself. One example is the sequence of chromosome 21. Other examples include chromosome 18, 13, X and Y. Yet other examples include mutated genetic sequences or genetic polymorphisms or copy number variations that a fetus may inherit from one or both of its parents. Yet other examples include sequences which are mutated, deleted, or amplified in a malignant tumor, e.g. sequences in which loss of heterozygosity or gene duplication occur. In some embodiments, multiple clinically relevant nucleic acid sequences, or equivalently multiple makers of the clinically relevant nucleic acid sequence, can be used to provide data for detecting the imbalance. For instance, data from five non-consecutive sequences on chromosome 21 can be used in an additive fashion for the determination of a possible chromosomal 21 imbalance, effectively reducing the need of sample volume to 1/5.

[0060] The term "*background nucleic acid sequence*" as used herein refers to a nucleic acid sequence whose normal ratio to the clinically relevant nucleic acid sequence is known, for instance a 1-to-1 ratio. As one example, the background nucleic acid sequence and the clinically relevant nucleic acid sequence are two alleles from the same chromosome that are distinct due to heterozygosity. In another example, the background nucleic acid sequence is one allele that is heterozygous to another allele that is the clinically relevant nucleic acid sequence. Moreover, some of each of the background nucleic acid sequence and the clinically relevant nucleic acid sequence may come from different individuals.

[0061] The term "*reference nucleic acid sequence*" as used herein refers to a nucleic acid sequence whose average concentration per reaction is known or equivalently has been measured.

[0062] The term "*overrepresented nucleic acid sequence*" as used herein refers to the nucleic acid sequence among two sequences of interest (e.g., a clinically relevant sequence and a background sequence) that is in more abundance than the other sequence in a biological sample.

[0063] The term "*based on*" as used herein means "*based at least in part on*" and refers to one value (or result) being used in the determination of another value, such as occurs in the relationship of an input of a method and the output of that method. The term "*derive*" as used herein also refers to the relationship of an input of a method and the output of that method, such as occurs when the derivation is the calculation of a formula.

[0064] The term "*quantitative data*" as used herein means data that are obtained from one or more reactions and that provide one or more numerical values. For example, the number of wells that show a fluorescent marker for a particular sequence would be quantitative data.

[0065] The term "*parameter*" as used herein means a numerical value that characterizes a quantitative data set and/or a numerical relationship between quantitative data sets. For example, a ratio (or function of a ratio) between a first amount of a first nucleic acid sequence and a second amount of a second nucleic acid sequence is a parameter.

[0066] The term "*cutoff value*" as used herein means a numerical value whose value is used to arbitrate between two or more states (e.g. diseased and non-diseased) of classification for a biological sample. For example, if a parameter is greater than the cutoff value, a first classification of the quantitative data is made (e.g. diseased state); or if the parameter is less than the cutoff value, a different classification of the quantitative data is made (e.g. non-diseased state).

[0067] The term "*imbalance*" as used herein means any significant deviation as defined by at least one cutoff value in a quantity of the clinically relevant nucleic acid sequence from a reference quantity. For example, the reference quantity could be a ratio of 3/5, and thus an imbalance would occur if the measured ratio is 1:1.

DETAILED DESCRIPTION OF THE INVENTION

[0068] This invention provides methods, systems, and apparatus for determining whether an increase or decrease compared to a reference (e.g. non-diseased) quantity of a clinically-relevant nucleic acid sequence in relation to other non-clinically-relevant sequences (e.g., a chromosomal or allelic imbalance) exists within a biological sample. One or more cutoff values are chosen for determining whether a change compared to the reference quantity exists (i.e. an imbalance), for example, with regards to the ratio of amounts of two sequences (or sets of sequences). The change detected in the reference quantity may be any deviation (upwards or downwards) in the relation of the clinically-relevant nucleic acid sequence to the other non-clinically-relevant sequences. Thus, the reference state may be any ratio or other quantity (e.g. other than a 1-1 correspondence), and a measured state signifying a change may be any ratio or other quantity that differs from the reference quantity as determined by the one or more cutoff values.

[0069] The clinically relevant nucleic acid sequence and the background nucleic acid sequence may come from a first type of cells and from one or more second types of cells. For example, fetal nucleic acid sequences originating from fetal/placental cells are present in a biological sample, such as maternal plasma, which contains a background of maternal nucleic acid sequences originating from maternal cells. Thus, in one embodiment, the cutoff value is determined based at least in part on a percentage of the first type of cells in a biological sample. Note the percentage of fetal sequences in a sample may be determined by any fetal-derived loci and not limited to measuring the clinically-relevant nucleic acid sequences. In another embodiment, the cutoff value is determined at least in part on the percentage of tumor sequences in a biological sample, such as plasma, serum, saliva or urine, which contains a background of nucleic acid sequences derived from the non-malignant cells within the body.

[0070] In yet another embodiment, the cutoff value is determined based on an average concentration of a sequence in a plurality of reactions. In one aspect, the cutoff value is determined from a proportion of informative wells that are estimated to contain a particular nucleic acid sequence, where the proportion is determined based on the above-mentioned percentage and/or average concentration. The cutoff value may be determined using many different types of methods, such as SPRT, false discovery, confidence interval, receiver operating characteristic (ROC). This strategy further minimizes the amount of testing required before confident classification can be made. This is of particular relevance to

plasma nucleic acid analysis where the template amount is often limiting. Although presented with respect to digital PCR, other methods may be used.

[0071] Digital PCR involves multiple PCR analyses on extremely dilute nucleic acids such that most positive amplifications reflect the signal from a single template molecule. Digital PCR thereby permits the counting of individual template molecules. The proportion of positive amplifications among the total number of PCRs analyzed allows an estimation of the template concentration in the original or non-diluted sample. This technique has been proposed to allow the detection of a variety of genetic phenomena (Vogelstein, B et al. 1999, supra) and has previously been used for the detection of loss of heterozygosity in tumor samples (Zhou, W. *et al.* 2002, supra) and in the plasma of cancer patients (Chang, HW *et al.* 2002, supra). Since template molecule quantification by digital PCR does not rely on dose-response relationships between reporter dyes and nucleic acid concentrations, its analytical precision should theoretically be superior to that of real-time PCR. Hence, digital PCR could potentially allow the discrimination of finer degrees of quantitative differences between target and reference loci.

[0072] To test this, we first assessed if digital PCR could determine the allelic ratio of *PLAC4* mRNA (Lo, YMD, *et al.* 2007 *Nat Med* 13, 218-223), a placental transcript from chromosome 21, in maternal plasma and thereby distinguish trisomy 21 and euploid fetuses. This approach is referred as the digital RNA-SNP method. We then evaluated whether the increased precision of digital PCR would allow the detection of fetal chromosomal aneuploidies without depending on genetic polymorphisms. We call this digital relative chromosome dosage (RCD) analysis. The former approach is polymorphism-dependent but requires less precision in quantitative discrimination while the latter approach is polymorphism-independent but requires a higher precision for quantitative discrimination.

I. DIGITAL RNA-SNP

A. Overview

[0073] Digital PCR is capable of detecting the presence of allelic ratio skewing of two alleles in a DNA sample. For example, it has been used to detect loss of heterozygosity (LOH) in a tumor DNA sample. Assuming that there are two alleles in the DNA sample, namely A and G, and the A allele would be lost in the cells with LOH. When LOH is present

in 50% of cells in the tumor sample, the allelic ratio of G:A in the DNA sample would be 2:1. However, if LOH is not present in the tumor sample, the allelic ratio of G:A would be 1:1.

[0074] FIG. 1 is a flowchart 100 illustrating a digital PCR experiment. In step 110, the DNA sample is diluted and then distributed to separate wells. Note that the inventors have determined that some plasma nucleic acid species are already quite diluted in the original sample. Accordingly, there is no need for dilution for some templates, if they are already present at the necessary concentrations. In the previous studies (e.g. Zhou et al 2002, supra), a DNA sample is diluted to an extent such that the average concentration of a specific "template DNA" is approximately 0.5 molecule of one of the two templates per well. Note that the term "template DNA" appears to refer to either the A or the G alleles, and that there is no rationale provided for this specific concentration.

[0075] In step 120, in each well, a PCR process is carried out to detect the A and/or the G allele simultaneously. In step 130, the markers in each well are identified (e.g. via fluorescence), e.g. A, G, A and G, or neither. In the absence of LOH, the abundance of the A and the G alleles in the DNA sample would be the same (one copy each per cell). Therefore, the probabilities of a well being positive for the A allele and for the G allele would be the same. This would be reflected by the similar numbers of wells being positive for the A or the G alleles. However, when LOH is present in 50% or greater of cells in a tumor sample, the allelic ratio of the G and the A alleles would be at least 2:1. Previous methods simply assumed that the sample was at least 50% cancerous. Thus, the probability of a well being positive for the G allele would be higher than that for the A allele. As a result, the number of wells being positive for the G allele would be higher than that for the A allele.

[0076] In step 140, to classify the digital PCR results, the number of wells being positive for each allele but not the other would be counted. In the above example, the number of wells being positive for the A allele but negative for the G allele, and the number of wells positive for the G allele but negative for the A allele are counted. In one embodiment, the allele showing less positive wells is regarded as the reference allele.

[0077] In step 150, the total number of informative wells is determined as the sum of the numbers of positive wells for either of the two alleles. In step 160, the proportion (P_r) of informative wells (an example of a parameter) contributed by the allele with more positive wells is calculated.

P_r = No. of wells only positive for the allele with more positive wells / Total no. of wells positive for only one allele (A or G).

Other embodiments could use all wells with one of the alleles divided by all wells with at least one allele.

[0078] In step 170, it is determined whether the value of P_r shows an allelic imbalance. As accuracy and efficiency are desired, this task is not straightforward. One method for determining an imbalance uses a Bayesian-type likelihood method, sequential probability ratio testing (SPRT). SPRT is a method which allows two probabilistic hypotheses to be compared as data accumulate. In other words, it is a statistical method to classify the results of digital PCR as being suggestive of the presence or absence of allelic skewing. It has the advantage of minimizing the number of wells to be analyzed to achieve a given statistical power and accuracy.

[0079] In an exemplary SPRT analysis, the experimental results would be tested against the null and alternative hypotheses. The alternative hypothesis is accepted when there is allelic ratio skewing in the sample. The null hypothesis is accepted when there is no allelic ratio skewing in the sample. The value P_r would be compared with two cutoff values to accept the null or alternative hypotheses. If neither hypothesis is accepted, the sample would be marked as unclassified which means that the observed digital PCR result is not sufficient to classify the sample with the desired statistical confidence.

[0080] The cutoff values for accepting the null or alternative hypotheses have typically been calculated based on a fixed value of P_r under the assumptions stated in the hypotheses. In the null hypothesis, the sample is assumed to exhibit no allelic ratio skewing. Therefore, the probabilities of each well being positive for the A and the G alleles would be the same and, hence, the expected value of P_r would be 1/2. In the alternative hypothesis, the expected value of P_r has been taken to be 2/3 or about halfway between 0.5 and 2/3, e.g. 0.585. Also, due to a limited number of experiments, one can choose an upper bound (.585+3/N) and a lower bound taken as (.585-3/N).

B. Detection of Down Syndrome

[0081] In one embodiment of the present invention, digital SNP is used to detect fetal Down syndrome from a pregnant woman's plasma. Using markers specific to fetal/placental cells, the ratio of alleles in chromosome 21 may be measured. For example, to determine if

an observed degree of overrepresentation of a *PLAC4* allele is statistically significant, SPRT is used.

[0082] According to one exemplary embodiment, digital RNA-SNP determines an imbalance in the ratio of polymorphic alleles of an A/G SNP, rs8130833, located on *PLAC4* mRNA which is transcribed from chromosome 21 and expressed by the placenta. For a heterozygous euploid fetus, the A and G alleles should be equally represented in the fetal genome (1:1 genomic ratio); while in trisomy 21, the trisomic chromosome 21 would be associated with an additional copy of one of the SNP alleles in the fetal genome giving a 2:1 ratio. The aim of the digital PCR analysis is to determine whether the amounts of the two *PLAC4* alleles in the analyzed sample are equal or otherwise. Thus, both the A and G *PLAC4* alleles are the target templates. A real-time PCR assay was designed to amplify *PLAC4* mRNA and the two SNP alleles were discriminated by TaqMan fluorescent probes. A schematic illustration of the analytical steps is shown in Fig. 2A.

[0083] FIG. 2A illustrates a digital RNA-SNP method 200 according to an embodiment of the present invention. In step 210, the sample is received. In step 220, the nucleic acid sequence, e.g. *PLAC4* mRNA, is quantified in extracted RNA samples. In one embodiment, this is done by real-time PCR for *PLAC4* mRNA. In one aspect, this step provides the operator with an idea about how much dilution is required before the target reaches the 'realm' of digital PCR analysis.

[0084] In step 230, the sample is diluted. In step 240, a concentration of the diluted sample is measured. The diluted sample concentration may be confirmed to be ~1 template/well (i.e. reference or non-reference sequence or either allele). Some embodiments use techniques described in section IV for this measurement. For example, we distributed the diluted sample to 96 wells for real-time PCR analysis to confirm that a usable dilution has been achieved. The dilution concentration may also be left as an unknown, thus removing this step as will be explained later.

[0085] In step 250, digital PCR is performed on each well of the array. For example, the same diluted sample was distributed to 384-wells for real-time PCR analysis. From the PCR results, an amount of markers for each nucleic acid sequence and the number of informative wells is identified. An informative well is defined as one that is only positive for the A or G allele but not both. In step 260, a calculation of an expected value of P_r is performed. These steps will be described in more detail later. The calculation includes determining a parameter

from values determined in step 250. For example, the actual average template concentration per well may be calculated.

[0086] In step 270, an SPRT or other likelihood-ratio test may be performed to determine whether or not an imbalance exists. For a euploid case, we expect an equal number of A-positive and G-positive wells. However, when template molecules from a trisomy 21 fetus are analyzed, the number of wells containing just one allele should be higher than that containing just the other allele. In short, allelic imbalance is expected for trisomy 21.

[0087] As mentioned above, SPRT is a Bayesian-type likelihood method which allows two probabilistic hypotheses to be compared as data accumulate. In digital PCR analysis for trisomy 21 detection, the alternative hypothesis is accepted when allelic imbalance exists (i.e. trisomy 21 detected); and the null hypothesis is accepted when there is no allelic imbalance (i.e. trisomy 21 not detected). The allele with the higher number of counts is referred as the potentially overrepresented allele and its proportion among all informative wells (P_r) would be calculated. SPRT is applied to determine if the P_r indicates sufficient degree of allelic imbalance that would be expected for a trisomy 21 sample.

[0088] Operationally, SPRT can be applied and interpreted through the use of graphs with a pair of SPRT curves that are constructed to define the probabilistic boundaries for accepting or rejecting either hypothesis. FIG. 3 illustrates a graph having SPRT curves for determining Down syndrome according to an embodiment of the present invention. The SPRT curves plot the required proportion of informative wells positive for the potentially overrepresented allele P_r, (y-axis) for a given total number of informative wells (x-axis) when confident classification could be made. As depicted in Fig. 3, the upper curve sets the probabilistic boundaries for accepting the alternative hypothesis; while the lower curve sets the probabilistic boundaries for accepting the null hypothesis.

[0089] The experimentally derived P_r value would be compared with the expected value of P_r to accept or reject either hypothesis. If the null hypothesis was accepted, the samples were classified as having been obtained from pregnant women with euploid fetuses. If the alternative hypothesis was accepted, the samples were classified as having been obtained from pregnant women with trisomy 21 fetuses. Alternatively, either hypothesis could not be accepted if the P_r for the given number of informative counts has not yet reached the required level of statistical confidence for disease classification. These cases were deemed unclassifiable until more data were available. If disease classification is not possible,

additional 384-well plates may be performed until the aggregated data become classifiable by SPRT.

[0090] SPRT thus offers the advantage that a smaller amount of testing is required for a given level of confidence than other statistical methods. In practical terms, SPRT allows the acceptance or rejection of either of the hypotheses as soon as the required amount of data has been accumulated and thus minimizes unnecessary additional analyses. This feature is of particular relevance to the analysis of plasma nucleic acids which are generally present at low concentrations where the number of available template molecules is limiting. In addition to a strict classification, the classification may also include a percent accuracy. For example, a classification resulting from a comparison with a cutoff value may provide that a sample shows a likelihood of a nucleic acid sequence imbalance with a certain percentage, or equivalently that determined imbalance is accurate to a certain percentage or other value.

[0091] A similar approach could be applied to determine the genotype of a fetus with regard to either a mutation or genetic polymorphism, using fetal nucleic acids in maternal plasma or serum. One should recall that a fetus would inherit half of its genome from its mother. As an illustration, consider a particular genetic locus with two alleles, A and B. If the mother is a heterozygote with a genotype of AB, the fetus could theoretically have a genotype of AA, BB or AB. If the fetus has a genotype of AB, i.e., the same as that of the mother, then there will only be nucleic acids of the AB genotype (from both the mother and fetus) in maternal plasma. Thus, nucleic acid or allelic balance is seen in maternal plasma. On the other hand, if the fetus has a genotype of AA or BB, then there would be allelic imbalance with an overrepresentation of the A or the B allele, respectively, in maternal plasma. This consideration is also applicable to disease-causing mutations (e.g. those causing cystic fibrosis or beta-thalassemia or spinal muscular atrophy), in which case A could be considered as the wildtype allele and B could be considered as the mutant allele.

II. DIGITAL RCD

[0092] A disadvantage of digital RNA-SNP is that it can only be applied to cases heterozygous for the analyzed SNP. One improvement is that it would be ideal if a noninvasive test for detecting fetal trisomy 21 or other fetal chromosomal aneuploidies (e.g. trisomy 18, 13, and the sex chromosome aneuploidies) based on circulating fetal nucleic acid analysis were independent of the use of genetic polymorphisms. Thus, in one embodiment, chromosome dosage is determined by digital PCR analysis of a non-polymorphic

chromosome 21 locus relative to one located on a reference chromosome, namely chromosome 1 in this study. A change of the ratio of chromosome 21 to chromosome 1 from 2:2 in the genome of a euploid fetus is differentiated from a trisomy 21 case. In digital PCR analysis for trisomy 21 detection, the two hypotheses to be compared would be the null hypothesis that there is no chromosomal imbalance (i.e. trisomy 21 not detected) and the alternative hypothesis that a chromosomal imbalance exists (i.e. trisomy 21 detected).

[0093] This approach can be generalized to the other chromosomes involved in other chromosomal aneuploidies, e.g. chromosome 18 in trisomy 18, chromosome 13 in trisomy 13, chromosome X in Turner syndrome. In addition, apart from chromosome 1, other chromosomes not involved in the aneuploidies concerned can also be used as a reference chromosome. A similar approach can also be applied to the detection of cancer, by analyzing the change of ratio of a chromosome commonly deleted, in part, in cancer, to a reference chromosome. Examples of the former include chromosome 5q in colorectal cancer, chromosome 3p in lung cancer and chromosome 9p in nasopharyngeal carcinoma. FIG. 2B lists some common cancer-related chromosomal aberrations which result in sequence imbalance.

[0094] FIG. 2A also illustrates a digital RCD method 205 according to an embodiment of the present invention. In one embodiment for steps 220-230, extracted DNA is quantified, for example, via Nanodrop techniques and diluted to a concentration of approximately one target template from either chromosomes 21 or the normalizing chromosome (such as chromosome 1) per well. In one embodiment in step 240, the confirmation may be performed by analyzing the diluted DNA sample by the assay using the chromosome 1 probe only in a 96-well format to confirm if ~37% level of the wells were negative before proceeding to digital RCD analysis using both TaqMan probes in 384-well plates. The significance of the 37% will be described later in Section IV.

[0095] The testing of step 240 and results of step 250 may be done with a real-time PCR assay designed to amplify a paralogous sequence (Deutsch, S. et al. 2004 *J Med Genet* 41, 908-915) present on both chromosomes which are distinguished by paralogous sequence variations that are discriminated by a pair of TaqMan probes. In this context, an informative well is defined as one that is positive for either the chromosome 21 or chromosome 1 locus but not both. For a euploid fetus, the number of informative wells positive for either locus should be approximately equal. For a trisomy 21 fetus, there should be an overrepresentation

of wells positive for chromosome 21 than chromosome 1. The exact proportion of the overrepresentation is described in the following sections.

III. INCORPORATING PERCENTAGE OF FETAL SEQUENCES

[0096] A disadvantage of embodiments of methods 200 and 205 described above is that fetal specific markers are required. Accordingly, in one embodiment of the present invention, non-fetal specific markers are used. In order to use such non-fetal specific markers, embodiments of the present invention measure the fractional concentration of fetal DNA in the maternal plasma (i.e. the biological sample). With such information, a more useful value of P_r may be calculated as follows.

[0097] Even with the small fractional percentage of fetal DNA in maternal plasma, a trisomy 21 fetus would contribute an additional dose of chromosome 21 sequences per genome-equivalent (GE) of fetal DNA released into maternal plasma. For example, a maternal plasma sample from a euploid pregnancy containing 50 GE/ml of total DNA with 5 GE/ml DNA contributed by the fetus (i.e. 10% fetal DNA fractional concentration) should contain a total of 100 copies (90 maternal copies + 10 fetal copies) of chromosome 21 sequences per milliliter of maternal plasma. For a trisomy 21 pregnancy, each fetal GE would contribute three copies of chromosome 21, resulting in a total of 105 copies/ml (90 maternal copies + 15 fetal copies) of chromosome 21 sequences in maternal plasma. At 10% fetal DNA concentration, the amount of chromosome 21 derived sequences in the maternal plasma of a trisomic pregnancy would therefore be 1.05 times that of a euploid case. Thus, if an analytical approach could be developed to determine this small degree of quantitative difference, a polymorphism-independent test for noninvasive prenatal diagnosis of fetal trisomy 21 would be achieved.

[0098] Accordingly, the degree of overrepresentation would be dependent on the fractional fetal DNA concentration in the analyzed DNA sample. For example, when placental DNA is analyzed, the theoretical RCD ratio in the fetal genome should be 3:2, i.e. 1.5-fold difference. However, as described above, the theoretical RCD ratio would decrease to 1.05 when a maternal plasma sample containing 10% fetal DNA is analyzed. The experimentally derived P_r is calculated by dividing the number of wells positive only for the chromosome 21 locus by the total number of informative wells. The experimentally derived P_r is subjected to the SPRT analysis with the calculated P_r and the theoretical RCD ratio.

[0099] FIG. 4 shows a method 400 of determining a disease state using a percentage of fetal nucleic acids according to an embodiment of the present invention. In step 410, the fractional percentage of fetal material is measured. In one embodiment, the fractional percentage is determined by measuring the amount of a fetal-specific marker (e.g. Y-chromosome, genetic polymorphism markers (e.g. SNPs), placental epigenetic signatures) in relation to a non-fetal-specific marker (i.e. gene sequence present in both mother and fetus). The actual measurement could be done by real-time PCR, digital PCR, sequencing reactions (including massively parallel genomic sequencing) or any other quantitative methods. In one aspect, it is preferable not to use the gene target that could potentially be in allelic imbalance for this measurement.

[0100] In step 420, digital PCR or other measurement method is performed, including diluting the sample, placing the diluted sample in the wells, and measuring the reactions in each well. In step 430, the PCR results are used to identify markers of different reference nucleic acid sequences (such as chromosomes or alleles). In step 440, the actual proportion (P_r) of the overrepresented sequence is calculated. In step 450, the cutoff value for determining a disease state is calculated using the percentage of fetal material in the sample. In step 460, from the actual P_r and the cutoff value, it is determined whether an imbalance exists.

[0101] In one embodiment, the fractional percentage of reference nucleic acid sequences is incorporated in a digital RNA-SNP method. Thus, when investigating a LOH due to cancer cells, this can be done with tumor samples with less than 50% cancer cells. It also may be used on samples with greater than 50% cancer cells to obtain a more accurate P_r and thus reduce the number of false positives, which would lead to incorrect diagnoses. In another embodiment, the fetal nucleic acid percentage is incorporated in a digital PCR method to determine if a fetus has inherited a parental gene mutation (e.g. that causing cystic fibrosis or beta-thalassemia or spinal muscular atrophy) or polymorphism from maternal plasma nucleic acid analysis.

IV. INCORPORATING AVERAGE CONCENTRATION PER WELL

[0102] Another disadvantage of previous methods (e.g. Zhou, W. *et al.* 2002, supra) is that the average concentration of templates per well (m) is required to be 1 per well. Given that it is difficult to obtain an exact concentration, this can lead to inaccuracies. Furthermore, even with an exact concentration of 1 template per well, previous methods have ignored the

statistical distribution of the templates in a well. In the previous methods, i.e. the old algorithm, the expected value of P_r for accepting the alternative hypothesis is assumed to be the allelic ratio and, thus, is independent of the average concentration of the template DNA per well.

[0103] However, due to a natural statistical variation of the templates in the diluted sample, there will not be exactly 1 template per well. Embodiments of the present invention measure the average concentration of at least one of the sequences, which is then used to calculate the cutoff value, i.e. the expected P_r. In one aspect, this calculation involves a statistical distribution to determine a probability of a well containing the different nucleic acid sequences, which is then used to determine the expected P_r.

[0104] In one embodiment, the average concentration is taken of one reference nucleic acid sequence, which in one instance is nucleic acid sequence with the lower concentration in the DNA sample. In the case of a sample without an imbalance, the concentrations of the two sequences in the sample would be the same and either one can be regarded as the reference allele. In the case of a sample with, for example, LOH, the allele which is deleted in the cancer cells would be regarded as the reference allele. The average concentration of the reference allele would be denoted as m_r. In another embodiment, the sequence with the higher concentration may be taken as the reference sequence.

A. Digital-SNP. Example using SPRT and digital PCR

[0105] FIG. 5 shows a method 500 of determining a disease state using an average template concentration according to an embodiment of the present invention. In step 510, an amount of the different sequences are measured. This may be done for example by counting the makers in a digital PCR experiment as explained above. However, it may be done by other methods that do not involve an amplification step or that do not use a fluorescent marker, but could use other properties, such as physical properties like mass, specific optical properties or base-pairing properties.

[0106] In step 520, the actual proportion of the overrepresented sequence is determined. This may be done as described above by taking the number of wells showing only that sequence and dividing by the number of informative wells. In step 530, the average concentration of at least one of the sequences is measured (the reference sequence). In one embodiment, the reference sequence is the overrepresented sequence. In another

embodiment, the reference sequence is the underrepresented sequence. The measurement may be done by counting the number of wells negative for the reference sequence in the digital PCR experiment. The relationship between the proportion of negative wells and the average template concentration is described by the Poisson distribution as described in the next subsection.

[0107] In step 540, an expected amount of wells positive for the different sequences is calculated, for example, using a Poisson distribution. The expected amount may be as a probability of the sequence per well, an average sequence per well, the number of wells containing the sequence or any other suitable quantity. In step 550, the expected P_r is calculated from the expected amounts. In step 560, a cutoff value is calculated from the expected P_r, for example, by using SPRT. In step 570, a classification of the nucleic acid sequence imbalance is determined. Specific aspects of method 500 are now described.

1. Determining expected amount of sequences

[0108] Once the average concentration per well (reaction or reaction mixture) is known from step 530, the expected number of wells showing that sequence may be calculated in step 540. This amount may be expressed as a %, a fractional value, or an integer value. Using a specific example for illustration, assume the average concentration of the reference template per well (m_r) is 0.5 per well and the genotype of the trisomy 21 fetus at the *PLAC4* SNP, rs8130833, is AGG. Therefore, the reference template would be the A allele and the overrepresented template would be the G allele.

[0109] In one embodiment, a Poisson distribution is assumed for the distribution of the A allele among the reaction mixtures of the wells of the measurement procedure, such as digital PCR. In other embodiments, other distribution functions are used, such as binomial distribution.

[0110] The Poisson equation is: $P(n) = \frac{m^n e^{-m}}{n!}$ where, n=number of template molecules per well; P(n)=probability of n template molecules in a particular well; and m= average number of template molecules in one well in a particular digital PCR experiment.

[0111] Accordingly, the probability of any well not containing any molecule of the A allele at an average A-allele concentration of 0.5 would be:

$$P(0) = \frac{0.5^0 e^{-0.5}}{0!} = e^{-0.5} = 0.6065.$$

[0112] Hence, the probability of any well containing at least one molecule of the A allele would be: 1 – 0.6065 = 0.3935. Therefore, ~39% of the wells would be expected to contain at least one molecule of the A allele.

[0113] As for the non-reference nucleic acid sequence, for each cell of a trisomy 21 fetus, the genomic ratio of A to G would be 1:2. Assuming that the A to G ratio in the extracted RNA or DNA sample would remain unchanged, the average concentration of the G allele per well would be two times that of the A allele, i.e. 2 x 0.5 = 1.

[0114] Accordingly, the probability of any well not containing any molecule of the G allele at an average G-allele concentration of 1 would be:

$$P(0) = \frac{1^0 e^{-1}}{0!} = e^{-1} = 0.3679$$

[0115] Hence, the probability of any well containing at least one molecule of the G allele would be: 1 – 0.3679 = 0.6321. Therefore, ~63% of the wells would be expected to contain at least one molecule of the G allele.

2. Determining proportion of overrepresented sequence

[0116] After the expected amounts are calculated, the proportion of the overrepresented nucleic acid sequence may be determined. Assuming that the filling of the wells with the A allele and the G allele are independent, the probability of a well containing both alleles would be 0.3935 × 0.6321 = 0.2487. Therefore, ~25% of the wells would be expected to contain both alleles.

[0117] The proportion of wells expected to contain the A allele but not the G allele would be the number of wells containing at least one A allele deducted by the number of wells containing both the A and G alleles: 0.3935 – 0.2487 = 0.1448. Similarly, the proportion of wells expected to contain the G allele but not the A allele would be: 0.6321 – 0.2487 = 0.3834. An informative well is defined as a well being positive for either the A allele or the G allele but not both.

[0118] Hence, the expected ratio of wells containing the A allele relative to the G allele in a digital RNA-SNP analysis is 0.1448/0.3834. In other words, the proportion of wells positive

only for the G allele is 2.65 times that of wells positive only for the A allele. This is in contrast to the fetal genomic ratio where the overrepresented allele is 2 times that of the other allele.

[0119] For SPRT analysis, the proportion of the informative wells positive for the overrepresented allele (P_r) is calculated and interpreted using SPRT curves. In the current example, the proportion of informative wells would be: 0.1448 + 0.3834 = 0.5282. Hence, the expected Pr of a trisomy 21 case at m_r 0.5 is: 0.3834/0.5282 = 0.73.

[0120] As the average template concentration (*m*) is a key parameter in the Poisson equation, the P_r would vary with *m*. FIG. 6 shows a table 600 that tabulates the expected digital RNA-SNP allelic ratio and P_r of trisomy 21 samples for a range of template concentrations expressed as the average reference template concentration per well (m_r) according to an embodiment of the present invention. Table 600 shows the expected allelic ratio and proportion of informative wells positive for the overrepresented allele (P_r) for a series of average reference template concentrations per well (m_r).

[0121] The expected value of P_r varies with the average concentration of the reference allele per well (m_r) in a non-linear fashion. As shown in table 600, the expected value of P_r for accepting the alternative hypothesis would increase with m_r. As the expected value of P_r for accepting the null hypothesis is fixed at 0.5, the samples with and without allelic imbalance would separate more widely in terms of the value of P_r when m_r increases. Note that in other embodiments the value for accepting the null hypothesis may be other than 0.5. This might occur when the normal ratio is different than 1:1, e.g., 5:3, and thus an imbalance would occur when the ratio deviates from 5:3. The difference in the amounts of the two different nucleic acid sequences would then be determined on a case by case basis.

[0122] However, as previous methods (e.g., Zhou, W. *et al.* 2002, supra) used a fixed expected value of P_r for LOH samples, they underestimated the value of P_r for those samples with LOH (alternative hypothesis accepted). The degree of underestimation would increase with m_r. In other words, the higher the average concentration of the reference allele in the DNA sample, the more inaccurate the old methods would be. This underestimation of P_r for accepting the alternative hypothesis would lead to the inaccurate calculation of the cutoff values for accepting both the null and alternative hypotheses.

3. Calculating cutoff values based on expected P_r

[0123] For embodiments using SPRT, one may use the equations for calculating the upper and lower boundaries of the SPRT curves from El Karoui at al. (2006). Furthermore, the level of statistical confidence preferred for accepting the null or alternative hypothesis could be varied through adjusting the threshold likelihood ratio in the equations. In this study, a threshold likelihood ratio of 8 is used because this value had been shown to provide satisfactory performance to discriminate samples with and without allelic imbalance in the context of cancer detection. Thus, in one embodiment, the equations for calculating the upper and lower boundaries of the SPRT curves are:

Upper boundary = $[(\ln 8)/N - \ln \delta]/\ln \gamma$

Lower boundary = $[(\ln 1/8)/N - \ln \delta]/\ln \gamma$

where, $\delta = (1 - \theta_1)/(1 - \theta_0)$

$\gamma = -(\theta_1(1 - \theta_0) / \theta_0(1 - \theta_1)$

θ_0 = proportion of informative wells containing the non-reference allele if the null hypothesis is true
= 0.5 (see below)

θ_1 = proportion of informative wells containing the non-reference (i.e. overrepresented) allele if the alternative hypothesis is true

N = number of informative wells
= number of wells positive for either allele only

(ln is a mathematical symbol representing the natural logarithm, i.e. $\log_e$.)

[0124] For the determination of θ_0 for accepting the null hypothesis, the sample is assumed to be obtained from a pregnant woman carrying a euploid fetus. Under this assumption, the expected number of wells positive for either template would be 1:1, and thus the expected proportion of informative wells containing the non-reference allele would be 0.5.

[0125] For the determination of θ_1 for accepting the alternative hypothesis, the sample is assumed to be obtained from a pregnant woman carrying a trisomy 21 fetus. The calculations for the expected P_r of trisomy 21 cases for digital RNA-SNP analysis are detailed in Table 600. Hence, θ_1 for digital RNA-SNP analysis refers to the data shown in the last column of table 600.

4. Measurement of average concentration

[0126] The measurement of m_r may be performed through a variety of mechanisms as known or will be known to one skilled in the art. In one embodiment, the value of m_r is determined during the experimental process of digital PCR analysis. As the relationship between the value of m_r and the total number of wells being positive for the reference allele can be governed by a distribution (e.g. the Poisson distribution), m_r can be calculated from the number of wells being positive for the reference allele using this formula:

$m_r = -\ln(1 - \text{proportion of wells being positive for the reference allele})$

Note that ln is the natural logarithm, i.e., $\log_e$. This approach provides a direct and precise estimation of m_r in the DNA sample used for the digital PCR experiment.

[0127] This method may be used to achieve a desired concentration. For example, the extracted nucleic acids of a sample may be diluted to a specific concentration, such as one template molecule per reaction well, as done in step 240 of method 200. In an embodiment using the Poisson distribution, the expected proportion of wells with no template may be calculated as e^{-m}, where m is the average concentration of template molecules per well. For example, at an average concentration of one template molecule per well, the expected proportion of wells with no template molecule is given by e^{-1}, i.e., 0.37 (37%). The remaining 63% of wells will contain one or more template molecules. Typically, the number of positive and informative wells in a digital PCR run would then be counted. The definition of informative wells and the manner by which the digital PCR data are interpreted depends on the application.

[0128] In other embodiments, the average concentration per well, m_r, is measured by another quantification method, for example, quantitative real-time PCR, semi-quantitative competitive PCR, real-competitive PCR using mass spectrometric methods, etc.

B. Digital RCD

[0129] Digital RCD using the average concentration may be performed in a similar manner to the digital SNP method described above. The numbers of wells positive for the reference chromosome (non-chromosome 21) marker, the chromosome 21 marker and both markers can be determined by digital PCR. The average concentration of the reference marker per well (m_r) can be calculated from the total number of wells negative for the reference marker,

irrespective of the positivity of the chromosome 21 marker, according to the Poisson probability function as in the calculation of m_r for the digital SNP analysis.

[0130] SPRT analysis may then be used for classifying a plasma sample as being obtained from a pregnant woman carrying a euploid or a trisomy 21 fetus. The null hypothesis would be accepted when the fetus was euploid. In this scenario, the expected ratio for the wells positive for the reference marker and chromosome 21 marker would be 1:1 and, thus, the expected proportion of informative wells with positive signal for chromosome 21 marker would be 0.5. The alternative hypothesis would be accepted when the fetus was trisomic for chromosome 21. In this scenario, if the sample DNA was solely derived from the fetus, the average concentration of the chromosome 21 marker in each well would be 3/2 times the average concentration of the reference marker (m_r).

[0131] While digital RCD may be used to determine chromosome dosage through the detection of fetal-specific markers, e.g. epigenetic signatures of the placenta (Chim, SSC. *et al.* 2005 *Proc Natl Acad Sci U S A* 102, 14753-14758), an embodiment of the digital RCD analysis uses non-fetal-specific markers. Thus, an additional step of measuring the percentage of fetal material would occur when non-fetal specific markers are used. Therefore, the average concentration of the chromosome 21 marker per well would be dependent on the proportion of the fetal DNA in the sample and can be calculated using: m_r [(200% + fetal DNA percentage)/200%].

[0132] Again using a specific example for illustration, the average concentration of the reference template, chromosome 1, per well (m_r) is assumed to be 0.5 and 50% of the DNA is assumed to be derived from the fetus and 50% of the DNA in the sample is derived from the mother.

[0133] Accordingly, using the Poisson distribution, the probability of any well not containing any molecule of the chromosome 1 locus when its average concentration is 0.5 per well would be:

$$P(0) = \frac{0.5^0 e^{-0.5}}{0!} = e^{-0.5} = 0.6065$$

[0134] Hence, the probability of any well containing at least one molecule of the chromosome 1 locus would be: $1 - 0.6065 = 0.3935$. Therefore, ~39% of the wells would be expected to contain at least one molecule of the locus.

[0135] For each cell of this trisomy 21 fetus, the genomic ratio of chromosome 21 to chromosome 1 would be 3:2. The ratio between chromosome 21 and chromosome 1 in the DNA sample would be dependent on the fractional fetal DNA concentration (fetal DNA %) and would be: 3 × fetal DNA % + 2 (1 – fetal DNA %) : 2 × fetal DNA % + 2 × (1 – fetal DNA %). Thus, in this case when the fractional fetal DNA concentration is 50%, the ratio would be: (3 × 50% + 2 × 50%) / (2 × 50% + 2 × 50%) = 1.25. If the digital SNP method did not use fetal specific markers, such a calculation could also be used to calculate the average concentration of the non-reference sequence.

[0136] Hence, when the average concentration of the chromosome 1 locus per well is 0.5, the average concentration of the chromosome 21 locus per well is: 1.25 × 0.5 = 0.625. Accordingly, the probability of any well not containing any molecule of the chromosome 21 locus when its average concentration is 0.625 per well would be:

$$P(0) = \frac{0.625^0 e^{-0.625}}{0!} = e^{-0.625} = 0.5353$$

[0137] Hence, the probability of any well containing at least one molecule of the chromosome 21 locus would be: 1 – 0.5353 = 0.4647. Therefore, ~46% of the wells would be expected to contain at least one molecule of the locus. Assuming that the filling of the wells with either loci are independent, the probability of a well containing both loci would be 0.3935 × 0.4647 = 0.1829. Therefore, ~18% of the wells would be expected to contain both loci.

[0138] The proportion of wells expected to contain the chromosome 1 locus but not the chromosome 21 locus would be the number of wells containing at least one chromosome 1 locus deducted by the number of wells containing both loci: 0.3935 – 0.1829 = 0.2106. Similarly, the proportion of wells expected to contain the chromosome 21 locus but not both loci would be: 0.4647 – 0.1829 = 0.2818. An informative well is defined as a well positive for either the chromosome 1 locus or the chromosome 21 locus but not both.

[0139] Hence, the expected chromosome 21 to chromosome 1 ratio in digital RCD analysis is 0.2818/0.2106 = 1.34. In other words, the proportion of wells positive only for the chromosome 21 locus is 1.34 times that of wells positive only for the chromosome 1 locus. This is in contrast to the ratio of 1.25 in the DNA sample.

[0140] For SPRT analysis, the proportion of the informative wells positive for the chromosome 21 locus (P_r) would need to be calculated and interpreted using SPRT curves. In the current example, the proportion of informative wells would be: 0.2106 + 0.2818 = 0.4924. Hence, the expected Pr of a trisomy 21 case with 50% fetal DNA at m_r 0.5 is: 0.2818/0.4924 = 0.57.

[0141] As the average template concentration (*m*) is a key parameter in the Poisson equation, the P_r would vary with *m*. FIG. 7 shows a table 700 that tabulates the expected P_r for the fractional fetal DNA concentrations of 10%, 25%, 50% and 100% in trisomy 21 samples at a range of template concentrations expressed as the average reference template concentration per well (m_r) according to an embodiment of the present invention. The calculations for the expected P_r of trisomy 21 cases for digital RCD analyses are detailed in table 700. Hence, θ_1 for digital RCD analysis of samples with varying fetal DNA fractional concentrations can be obtained from the columns showing the corresponding expected P_r-values in table 700.

C. Results

1. Comparison of different m_r

[0142] The basis for the difference between the theoretical (as in the fetal genome) and experimentally-expected degree of allelic or chromosomal imbalance and the calculations to determine the latter for a range of m_r values are shown in tables 600 and 700. In digital RNA-SNP analysis of a trisomy 21 sample, when m_r=0.5, wells containing just the overrepresented allele with respect to wells containing just the reference allele, namely the digital RNA-SNP ratio, is 2.65 (table 600). In digital RCD analysis of a specimen composed of 100% fetal DNA, when m_r=0.5, wells positive solely for the chromosome 21 locus with respect to those positive solely for the chromosome 1 locus, namely the digital RCD ratio, is 1.7 (table 700) (P_r = 0.63, thus the digital RCD ratio is 0.63/(1-0.63) = 1.7). As the fractional fetal DNA concentration decreases, the digital RCD ratio decreases for the same m_r (table 700).

[0143] As shown in tables 600 and 700, the extent of allelic or chromosomal overrepresentation increases with m_r. However, the percentage of informative wells approaches its maximum around m_r=0.5 and decreases gradually with further increase in m_r. In practice, the decline in the proportion of informative wells could be compensated by increasing the total number of wells analyzed if the amount of template molecule for the

specimen is not limiting, but additional wells would require an increase in reagent costs. Hence, optimal digital PCR performance is a trade-off between the template concentration and total number of wells tested per sample.

2. Example using SPRT curves

[0144] As discussed above, the expected degree of allelic or chromosomal imbalance for a digital PCR experiment is dependent on the actual template concentration per reaction mixture (e.g. a well). We describe the template concentration based on the reference allele, i.e. the average reference template concentration per well (m_r). As shown in the above equation, the expected P_r can be used to determine the plotting of the upper and the lower SPRT curves. Since the expected P_r is in turn dependent on the value of m_r, the plotting of the SPRT curves would essentially be dependent on the value of m_r. Thus, in practice, a set of SPRT curves relevant for the actual m_r of a digital PCR dataset would need to be used for the interpretation of the P_r from that particular run.

[0145] FIG. 8 shows a plot 800 illustrating the degree of differences in the SPRT curves for m_r values of 0.1, 0.5 and 1.0 for digital RNA-SNP analysis according to an embodiment of the present invention. Each set of digital PCR data should be interpreted with the specific curves relevant to the exact m_r value of that particular run. Note that since the expected degree of allelic or chromosomal imbalance for the digital RNA-SNP and RCD approaches are different (2:1 for the former and 3:2 for the latter), different sets of SPRT curves are needed for the two digital PCR systems. The experimentally derived P_r is interpreted with the relevant SPRT curves selected by the corresponding m_r of the digital PCR run. This is in contrast to the previous reported use of SPRT for molecular detection of LOH by digital PCR where a fixed set of curves had been used.

[0146] The practical manner for interpreting the digital PCR data using SPRT is illustrated below using a hypothetical digital RNA-SNP run. After digital RNA-SNP analysis of each case, the number of wells positive for the A allele only, the G allele only or both alleles are counted. The reference allele is defined as the allele with the smaller number of positive wells. The value of m_r is calculated using the total number of wells negative for the reference allele, irrespective of whether the other allele is positive, according to the Poisson probability density function. The data of our hypothetical example are as follows:

[0147] In a 96-well reaction, 20 wells are positive for the A allele only, 24 wells are positive for the G allele only, and 33 wells are positive for both alleles. The A allele is regarded as the reference allele because there are less A-positive than G-positive wells. The number of wells negative for the reference allele is 96 – 20 – 33 = 43. Therefore, m_r can be calculated using the Poisson equation: $-\ln(43/96) = 0.80$. The experimentally determined P_r of this case is: 24 / (20 + 24) = 0.55.

[0148] According to table 600, the expected P_r of a trisomy 21 sample at $m_r = 0.8$ is 0.76. Thus, θ_1 is 0.76 for this case. The SPRT curves based on $\theta_1 = 0.76$ would be used to interpret the experimentally derived P_r of this case which is 0.55. When P_r=0.55 is fitted onto the relevant SPRT curves, the data point falls under the lower curve. Hence, the case is classified as euploid, see FIG. 3.

3. Comparison to old method

[0149] FIG. 9A shows a table 900 of a comparison of the effectiveness of the new and old SPRT algorithms for classifying euploid and trisomy 21 cases in 96-well digital RNA-SNP analyses. FIG. 9B shows a table 950 of a comparison of the effectiveness of the new and old SPRT algorithms for classifying euploid and trisomy 21 cases in 384-well digital RNA-SNP analyses. The new algorithm refers to the selection of SPRT curves specific for the m_r derived from the digital PCR data. The old algorithm refers to the use of a fixed set of SPRT curves for all digital PCR runs. The effect of incorrect calculation of the cutoff values on the accuracy of classification is revealed by the simulation analysis shown in table 900.

[0150] Compared with the use of a fixed set of SPRT curves in previous studies, the proportion of unclassifiable data is much lower with our approach, as shown in tables 900 and 950. For example, using our approach, at m_r=0.5, 14% and 0% of trisomy 21 samples would be unclassifiable for 96-well and 384-well digital RNA-SNP analyses, respectively, but 62% and 10%, respectively, with the use of fixed curves (Table 900). Hence, our approach allows disease classification with lesser number of informative wells.

[0151] As shown in table 900, the new algorithm is more accurate in classifying the samples as having or not having allelic ratio skewing for all values of m_r from 0.1 to 2.0. For example, when m_r equals 1.0 and a 96–well digital RNA-SNP run is performed, the new algorithm correctly classifies 88% and 92% of samples with and without allelic ratio

skewing, respectively, whereas the percentage of correct classification of samples with and without allelic ratio skewing using the old algorithm is only 19% and 36%, respectively.

[0152] Using the new algorithm, the separation of samples with and without allelic ratio skewing would increase with m_r. As a result, the classification accuracies would increase with m_r. The effect of increase in separation of the two groups of samples on the classification accuracy would diminish when m_r increases to beyond 2.0 because of the reduction in the percentage of informative wells. In contrast, using the old algorithm, the classification accuracies significantly reduce when m_r increases because of the increased deviation of expected P value from its true value.

[0153] Our experimental and simulation data show that digital RNA-SNP is an effective and accurate method for trisomy 21 detection. As *PLAC4* mRNA in maternal plasma is derived purely from the fetus, for 12 of the 13 maternal plasma samples tested, only one 384-well digital PCR experiment was required for correct classification (Table 1350 of FIG. 13B). This homogenous, real-time digital PCR-based approach thus offers an alternative to the mass spectrometry-based approach for RNA-SNP analysis (Lo, YMD, *et al.* 2007 *Nat Med*, supra). Apart from placental-specific mRNA transcripts, we also envision that other types of fetal-specific nucleic acid species in maternal plasma could be used for digital PCR-based detection of fetal chromosomal aneuploidies. One example is fetal epigenetic markers (Chim, SSC et al. (2005) *Proc Natl Acad Sci U S A* 102, 14753-14758; Chan, KCA *et al.* (2006) *Clin Chem* 52, 2211-2218), which have recently been used for the noninvasive prenatal detection of trisomy 18 using the epigenetic allelic ratio (EAR) approach (Tong, YK et al. (2006) *Clin Chem* 52, 2194-2202). Thus, we predict that digital EAR would be a possible analytical technique.

V. INCREASING %, MULTIPLE MARKERS, AND PCR ALTERNATIVES

[0154] As described above, the application of embodiments of the present invention to DNA extracted from maternal plasma can be complicated when the fetal DNA only constitutes a minor fraction of maternal plasma DNA, with a mean fractional concentration of some 3% between weeks 11 and 17 of gestation. Nevertheless, as shown herein, digital RCD allows aneuploidy detection even when the aneuploid DNA is present as a minor population. With a decreasing fractional concentration of fetal DNA, such as may be present during early gestation, a larger number of informative counts is needed for digital RCD. A significance of the present work, as summarized in table 1200 of FIG. 12, is that we have provided a set of

benchmark parameters, e.g. fractional fetal DNA and total template molecules required, which diagnostic assays can be built upon. In our opinion, 7680 total number of reactions for a fractional fetal DNA concentration of 25% is a particularly attractive set of benchmark parameters. These parameters should allow euploid and trisomy 21 samples to be classifiable correctly 97% of the time, as shown in table 1200.

[0155] The number of plasma DNA molecules that are present per unit volume of maternal plasma is limited (Lo, YMD. *et al.* 1998 *Am J Hum Genet* 62, 768-7758). For example, in early pregnancy, the median maternal plasma concentration of an autosomal locus, the *β-globin* gene, has been shown to be 986 copies/mL, with contributions from both the fetus and mother (Lo, YMD. *et al.* 1998 *Am J Hum Genet* 62, 768-7758). To capture 7,680 molecules, DNA extracted from some 8 mL of maternal plasma would be needed. This volume of plasma, obtainable from ~15 mL of maternal blood, is at the limit of routine practice. However, we envision that multiple sets of chr21 and reference chromosome targets can be combined for digital RCD analysis. For 5 pairs of chr21 and reference chromosome targets, just 1.6 mL of maternal plasma would be needed to provide the number of template molecules needed for analysis. Multiplex single molecule PCR could be performed. The robustness of such multiplex single molecule analysis has been demonstrated previously for single molecule haplotyping (Ding, C. and Cantor, CR. 2003 *Proc Natl Acad Sci U S A* 100, 7449-7453).

[0156] Alternatively, to achieve a fractional fetal DNA concentration of 25%, methods may allow the selective enrichment of fetal DNA (Li, Y. et al. 2004 *Clin Chem* 50, 1002-1011) or the suppression of the maternal DNA background (Dhallan, R et al. 2004 *JAMA* 291, 1114-1119) or both, in maternal plasma. Apart from such physical methods for fetal DNA enrichment and maternal DNA suppression, it would also be possible to use a molecular enrichment strategy, such as by targeting fetal DNA molecules which exhibit a particular DNA methylation pattern (Chim, SSC et al, 2005 *Proc Natl Acad Sci USA* 102,14753-14758, Chan, KCA et al. 2006 *Clin Chem* 52, 2211-2218; Chiu, RWK *et al.* 2007 *Am J Pathol* 170, 941-950.)

[0157] Additionally, there are now a number of alternative approaches to the manual set up of digital real-time PCR analyses as used in the current study for conducting digital PCR. These alternative approaches include microfluidics digital PCR chips (Warren, L et al. 2006 *Proc Natl Acad Sci U S A* 103, 17807-17812; Ottesen, EA et al. 2006 *Science* 314, 1464-

1467), emulsion PCR (Dressman, D et al. 2003 *Proc Natl Acad Sci U S A* 100, 8817-8822), and massively parallel genomic sequencing (Margulies, M. *et al.* 2005 *Nature* 437, 376-380) using for example the Roche 454 platform, the Illumina Solexa platform, and the SOLiD™ system of Applied Biosystems, etc. With regard to the latter, our method is also applicable to massively parallel sequencing methods on single DNA molecules, which do not require an amplification step, such as the Helicos True Single Molecule DNA sequencing technology (Harris TD et al. 2008 Science, 320, 106-109), the single molecule, real-time (SMRT™) technology of Pacific Biosciences, and nanopore sequencing (Soni GV and Meller A. 2007 *Clin Chem* 53, 1996-2001). With the use of these methods, digital RNA-SNP and digital RCD could be performed rapidly on a large number of samples, thus enhancing the clinical feasibility of the methods proposed here for noninvasive prenatal diagnosis.

EXAMPLES

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

I. COMPUTER SIMULATIONS

[0158] Computer simulation was performed to estimate the accuracy of diagnosing trisomy 21 using the SPRT approach. The computer simulation was performed with the Microsoft Excel 2003 software (Microsoft Corp., USA) and SAS 9.1 for Windows software (SAS Institute Inc., NC, USA). The performance of digital PCR is an interplay between the reference template concentration (m_r), number of informative counts and projected degree of allelic or chromosomal imbalance (P_r). Separate simulations were performed for a range of each of these variables. Since the decision boundaries of the SPRT curves for digital RNA-SNP and digital RCD were different, the simulation analyses for the two systems were performed separately.

[0159] For each digital PCR condition simulated (i.e. m_r, fetal DNA fractional concentration, total well number), two rounds of simulation were performed. The first round simulated the scenario that the tested samples were obtained from pregnant women carrying euploid fetuses. The second round simulated the scenario when the tested samples were obtained from pregnant women carrying trisomy 21 fetuses. For each round, testing of 5000 fetuses was simulated.

A. RNA-SNP

[0160] For digital RNA-SNP, simulations of a 384-well experiment with m_r=0.1 to m_r=2.0 were performed. At each m_r value, we simulated the scenario whereby 5000 euploid and 5000 trisomy 21 fetuses were tested. The SPRT curves appropriate for the given m_r were used to classify the 10,000 fetuses. FIG. 10 is a table 1000 showing the percentages of fetuses correctly and incorrectly classified as euploid or aneuploid and those not classifiable for the given informative counts according to an embodiment of the present invention. The accuracies for diagnosing euploid and aneuploid cases are both 100%, for m_r between 0.5 and 2.0. When m_r=0.1, only 57% and 88% of euploid and trisomy 21 fetuses could be accurately classified after the analysis of 384 wells.

[0161] The simulation data were generated as described in the following steps:

[0162] In step 1, for each well, two random numbers were generated using the Random(Poisson) function of the SAS program (www.sas.com/technologies/analytics/statistics/index.html) to represent the A and the G alleles, respectively. The Random(Poisson) function would generate positive integers starting from 0 (i.e. 0, 1, 2, 3,....) and the probability of each integer being generated was dependent on the probability of this number according to the Poisson probability density function for a given mean value which represented the average concentration of the alleles per well. A well was regarded as positive for the A allele if the random number representing the A allele was larger than zero, i.e. contained 1 or more molecules of the A allele. Similarly, the well was regarded as positive for the G allele if the random number representing the G allele was larger than zero.

[0163] To simulate the scenario of a pregnant woman carrying a euploid fetus, the same mean value was used for generating the random numbers for the A allele and the G allele. For example, in the analysis simulating digital RNA-SNP analyses at m_r = 0.5, the mean value for either the A allele of the G allele was set identically to 0.5 which meant an average concentration for either allele of 0.5 molecule per well. Using the Poisson equation, at a mean concentration of 0.5, the proportion of wells being positive for the A or the G alleles would be the same and was 0.3935, see table 600.

[0164] When simulating the digital RNA-SNP analysis of a pregnant woman with a trisomy 21 fetus at m_r = 0.5, the average concentration of the overrepresented allele per well

would be expected to be 2 times of that of the reference allele, i.e. 1. In this situation, the probability of a well being positive for the overrepresented allele was 0.6321, see table 600.

[0165] After generating a random number for a digital PCR well, the well could be classified as one of the following statuses:

a. negative for both the A and the G alleles
b. positive for both the A and the G alleles
c. positive for the A allele but negative for the G allele
d. positive for the G allele but negative for the A allele

[0166] In step 2, step 1 was repeated until the desired number of wells, 384 wells for the current simulation, had been generated. The numbers of wells positive for the A allele only and the G allele only were counted. The allele with less positive wells was regarded as the reference allele and the allele with more positive wells was regarded as the potentially overrepresented allele. The number of the informative wells was the total number of wells positive for either allele but not both. The proportion of informative wells containing the potentially overrepresented allele (P_r) was then calculated. The upper and lower boundaries for the relevant SPRT curves to accept the null or alternative hypothesis were calculated according to an embodiment of the present invention.

[0167] In step 3, 5000 simulations were performed for each of the two scenarios of the pregnant woman carrying a euploid or a trisomy 21 fetus. Each simulation could be regarded as an independent biological sample obtained from pregnant women. In Table 1000, the correct classification of euploid cases refers to those euploid cases in which the null hypothesis was accepted and the incorrect classification of euploid cases refers to those euploid cases in which the alternative hypothesis was accepted. Similarly, those trisomy 21 cases in which the alternative hypothesis was accepted were regarded as correctly classified and those trisomy 21 cases in which the null hypothesis was accepted were regarded as incorrectly classified. For both groups, those cases in which neither the null or alternative hypothesis was accepted after the pre-specified total number of wells had been simulated were regarded as unclassified.

[0168] In step 4, steps 1 to 3 were performed for m_r ranging from 0.1 to 2.0 at increments of 0.1.

B. RCD

[0169] FIG. 11 is a table 1100 showing computer simulations for digital RCD analysis for a pure (100%) fetal DNA sample for m_r ranging from 0.1 to 2.0 according to an embodiment of the present invention. As the fractional fetal DNA concentration becomes lower, the degree of chromosome 21 overrepresentation diminishes and thus a larger number of informative wells for accurate disease classification is required. Hence, simulations were further performed for fetal DNA concentrations of 50%, 25% and 10% for a total well number ranging from 384 to 7680 wells at m_r=0.5.

[0170] FIG. 12 is a table 1200 showing results of computer simulation of accuracies of digital RCD analysis at m_r=0.5 for the classification of samples from euploid or trisomy 21 fetuses with different fractional concentrations of fetal DNA according to an embodiment of the present invention. The performance of digital RCD is better for cases with a higher fetal DNA fractional concentration. At 25% fetal DNA concentration and a total number of 7680 PCR analyses, 97% of both euploid and aneuploid cases would be classifiable with no incorrect classification. The remaining 3% of cases require further analyses until classification can be achieved.

[0171] The procedures for simulating digital RCD analyses were similar to those described for digital RNA-SNP analysis. The steps for the simulations are described below:

[0172] In step 1, two random numbers under the Poisson probability density function were generated to represent the reference locus, chromosome 1, and the chromosome 21 locus. For subjects carrying euploid fetuses, the average concentrations of both the chromosome 1 and chromosome 21 loci were the same. In this simulation analysis, an average template concentration of 0.5 for each locus per well was used. For subjects carrying trisomy 21 fetuses, the m_r in this simulation was 0.5 but the average concentration of the chromosome 21 locus per well would depend on the fractional fetal DNA concentration in the tested sample, as shown in Table 700. The distribution of the reference and/or the chromosome 21 loci to a well was determined by the random numbers representing the respective locus which were generated according to the Poisson probability density function with the appropriate average concentration of the locus per well.

[0173] In step 2, step 1 was repeated until the desired number of wells had been generated, e.g. 384 wells for a 384-well plate experiment. The numbers of wells positive for chromosome 1 only and chromosome 21 only were counted. The number of the informative

wells was the total number of wells positive for either one of the chromosomes but not both. The proportion of informative wells positive for chromosome 21 (P_r) was then calculated. The upper and lower boundaries of the relevant SPRT curves to accept the null or alternative hypothesis were calculated as described in the earlier section on SPRT analysis.

[0174] In step 3, 5000 simulations were performed for each of the two scenarios of the pregnant woman carrying a euploid or a trisomy 21 fetus. Each simulation could be regarded as an independent biological sample obtained from pregnant women. In Table 1100, the correct classification of euploid cases refers to those euploid cases in which the null hypothesis was accepted and the incorrect classification of euploid cases refers to those euploid cases in which the alternative hypothesis was accepted. Similarly, those trisomy 21 cases in which the alternative hypothesis was accepted were regarded as correctly classified and those trisomy 21 cases in which the null hypothesis was accepted were regarded as incorrectly classified. For both groups, those cases in which neither the null or alternative hypothesis was accepted after the pre-specified total number of wells had been simulated were regarded as unclassified.

[0175] In step 4, steps 1 to 3 were repeated for samples with 10% 25%, 50% and 100% fetal DNA at total well numbers ranging from 384 to 7680.

II. VALIDATION OF TRISOMY 21 DETECTION

A. RNA-SNP for *PLAC4*

[0176] The practical feasibility of digital RNA-SNP was demonstrated using the rs8130833 SNP on the *PLAC4* gene on chromosome 21 (Lo, YMD et al. 2007 *Nat Med* 13, 218-223). Placental DNA and RNA samples from two euploid and two trisomy 21 heterozygous placentas were analyzed. The placental DNA samples were analyzed with the digital RNA-SNP protocol but with the omission of the reverse transcription step, thus essentially converting the procedure to digital *DNA*-SNP analysis. To strike a balance between the chance of correct case classification and the proportion of informative wells, we diluted the samples aiming for one allele of any type per well and confirmed by a 96-well digital PCR analysis. This was followed by a 384-well digital RNA-SNP experiment. P_r and m_r were calculated and the SPRT curve for this m_r value was used for data interpretation.

[0177] FIG. 13A shows a table 1300 of digital RNA-SNP analysis in placental tissues of euploid and trisomy 21 pregnancies according to an embodiment of the present invention.

Genotypes were determined by mass spectrometric assay. "Euploid" was assigned when the experimentally obtained P_r was below the unclassifiable region; "T21", representing trisomy 21, was assigned when the experimentally obtained P_r was above the unclassifiable region. T21, trisomy 21. Each of these cases was correctly classified, using both the DNA and RNA samples, with one 384-well experiment.

[0178] We further tested plasma RNA samples from nine women carrying euploid fetuses and four women carrying trisomy 21 fetuses. FIG. 13B shows a table 1350 of digital RNA-SNP analysis of maternal plasma from euploid and trisomy 21 pregnancies according to an embodiment of the present invention. All of the cases were correctly classified. Initial results for one trisomy 21 case (M2272P) fell within the unclassifiable region between the SPRT curves after one 384-well experiment. Thus, an additional 384-well run was performed. New m_r and P_r values were calculated from the aggregated data from the total of 768 wells and the classification was performed using a new set of SPRT curves selected based on this m_r value. The case was then scored correctly as aneuploid.

[0179] Our experimental and simulation data show that digital RNA-SNP is an effective and accurate method for trisomy 21 detection. As *PLAC4* mRNA in maternal plasma is derived purely from the fetus, for 12 of the 13 maternal plasma samples tested, only one 384-well digital PCR experiment was required for correct classification. This homogenous, real-time digital PCR-based approach thus offers an alternative to the mass spectrometry-based approach for RNA-SNP analysis. Apart from placental-specific mRNA transcripts, we also envision that other types of fetal-specific nucleic acid species in maternal plasma could be used for digital PCR-based detection of fetal chromosomal aneuploidies. One example is fetal epigenetic markers which have recently been used for the noninvasive prenatal detection of trisomy 18 using the epigenetic allelic ratio (EAR) approach (Tong YK et al. 2006 *Clin Chem,* 52, 2194-2202). Thus, we predict that digital EAR would be a possible analytical technique.

B. RCD

[0180] The practical feasibility of digital RCD for trisomy 21 detection was also investigated using a PCR assay targeting paralogous sequences on chromosome 21 and 1. Paralogous loci were used here by way of examples. Non-paralogous sequences on chromosome 21 and any other reference chromosome can also be used for RCD. Placental DNA samples from two euploid and two trisomy 21 placentas were diluted to approximately

one target template from either chromosome per well and confirmed by a 96-well digital PCR analysis. Each confirmed sample was analyzed by a 384-well digital RCD experiment and the values of P_r and m_r were calculated. For digital RCD, the chromosome 1 paralog was the reference template. This m_r value was used to select a corresponding set of SPRT curves for data interpretation. All of the placental samples were correctly classified as shown in FIG. 14A.

[0181] To demonstrate that the digital RCD approach can be used to detect trisomy 21 DNA which is mixed with an excess of euploid DNA, such as the scenario of fetal DNA in maternal plasma, mixtures containing 50% and 25% of trisomy 21 placental DNA in a background of euploid maternal blood cell DNA were analyzed. Placental DNA from 10 trisomy 21 and 10 euploid cases was each mixed with an equal amount of euploid maternal blood cell DNA, thus producing twenty 50% DNA mixtures. FIG. 14B shows a plot 1440 illustrating the SPRT interpretation for RCD analysis of the 50% fetal DNA mixtures according to an embodiment of the present invention. Similarly, placental DNA from 5 trisomy 21 and 5 euploid cases was each mixed with 3 times excess of euploid maternal blood cell DNA, thus producing ten 25% DNA mixtures. FIG. 14C shows a plot 1470 illustrating the SPRT interpretation for RCD analysis of the 25% fetal DNA mixtures. All the euploid and aneuploid DNA mixtures were correctly classified, as shown in FIGS. 14B and 14C.

[0182] Each sample reached the point of being classifiable after a number of 384-well digital PCR analyses as marked on FIGS. 14B and 14C. For the 50% DNA mixtures, the number of 384-well plates required ranged from one to five. For the 25% DNA mixtures, the number of 384-well plates required ranged form one to seven. The cumulative proportion of cases being correctly classified with the addition of every 384 digital PCR analyses were as predicted by the computer simulation presented in Table 1200.

III. METHOD WITH DIGITAL PCR

A. Digital RNA-SNP

[0183] All RNA samples were first reverse transcribed with a gene-specific reverse transcription primer using the ThermoScript reverse transcriptase (Invitrogen). Sequence of the reverse transcription primer was 5'-AGTATATAGAACCATGTTTAGGCCAGA-3' (Integrated DNA Technologies, Coralville, IA). The subsequent treatment of the reverse transcribed RNA (i.e. the cDNA) samples for digital RNA-SNP, and DNA samples (e.g.

placental DNA) was essentially the same. Prior to digital PCR analysis, DNA and the cDNA samples were first quantified using a real-time PCR assay towards *PLAC4*, consisting of primers 5′-CCGCTAGGGTGTCTTTTAAGC-3′, 5′-GTGTTGCAATACAAAATGAGTTTCT-3′, and the fluorescent probe 5′-(FAM)ATTGGAGCAAATTC(MGBNFQ)-3′ (Applied Biosystems, Foster City, CA), where FAM is 6-carboxyfluorescein and MGBNFQ is a minor groove binding non-fluorescent quencher.

[0184] A calibration curve was prepared by serial dilutions of HPLC-purified single-stranded synthetic DNA oligonucleotides (Proligo, Singapore) specifying the amplicon. The sequence was 5′-CGCCGCTAGGGTGTCTTTTAAGCTATTGGAGCAAATTCAAATTTGGCTTAAAGAAAAAGAAACTCATTTTGTATTGCAACACCAGGAGTATCCCAAGGGACTCG-3′. The reaction was set up using 2X TaqMan Universal PCR Master Mix (Applied Biosystems) in a reaction volume of 25 μL. 400 nM of each primer and 80 nM of the probe were used in each reaction. The reaction was initiated at 50°C for 2 min, followed by 95°C for 10 min and 45 cycles of 95°C for 15s and 60°C for 1 min in an ABI PRISM 7900HT Sequence Detection System (Applied Biosystems). Serial dilutions of the DNA or cDNA samples were then undertaken such that the subsequent digital PCR amplification could be performed at approximately one template molecule per well. At such a concentration, it was expected that approximately 37% of the reaction wells would show negative amplification and was first confirmed by conducting a 96-well digital real-time PCR analysis. This was followed by digital RNA-SNP analysis conducted in 384-well plates using a set of non-intron spanning primers: the forward primer 5′-TTTGTATTGCAACACCATTTGG-3′ and the gene-specific reverse transcription primer described above.

[0185] Two allele-specific TaqMan probes targeting each of the two alleles of the rs8130833 SNP on the *PLAC4* sequence were designed. Their sequences were 5′-(FAM)TCGTCGTCTAACTTG(MGBNFQ)-3′ and 5′-(VIC)ATTCGTCATCTAACTTG(MGBNFQ) for the G and A alleles, respectively. The reaction was set up using 2X TaqMan Universal PCR Master Mix in a reaction volume of 5 μL. Each reaction contains 1X TaqMan Universal PCR Master Mix, 572 nM of each primer, 107 nM of the allele-G-specific probe and 357 nM of the allele-A-specific probe. The reaction was carried out in the ABI PRISM 7900HT Sequence Detection System. The reaction was initiated at 50°C for 2 min, followed by 95°C for 10 min and 45 cycles of 95°C

for 15s and 57°C for 1 min. During the reaction, the fluorescence data were collected by the "Absolute Quantification" application of the SDS 2.2.2 software (Applied Biosystems). The software automatically calculated the baselines and the threshold values. The number of wells which were positive for either the A or the G alleles was recorded and subjected to SPRT analysis.

B. Digital RCD analysis

[0186] All placental and maternal buffy coat DNA samples used in this study were first quantified by the NanoDrop spectrophotometer (NanoDrop Technology, Wilmington, DE). The DNA concentration is converted to copies/μL using a conversion of 6.6 pg/cell. The amount of DNA corresponding to approximately one template per well was determined by serially diluting the DNA samples and confirmed with the real-time PCR assay in a 96-well format where we expect approximately 37% of the wells to show negative amplification. The PCR setup for the confirmatory plate was the same as described below except that only the probe for the reference chromosome was added. In the digital RCD analysis, the paralogous loci on chromosome 21 and 1 (Deutsch, S. *et al.* 2004 *J Med Genet* 41, 908-915) were first co-amplified by forward primer 5'-GTTGTTCTGCAAAAAACCTTCGA-3' and reverse primer 5'-CTTGGCCAGAAATACTTCATTACCATAT-3'. Two chromosome-specific TaqMan probes were designed to target the chromosome 21 and 1 paralogs, and their sequences were 5'-(FAM)TACCTCCAT AATGAGTAAA(MGBNFQ)-3' and 5'-(VIC)CGTACCTCTGTAATGTGTAA(MGBNFQ)-3', respectively. Each reaction contained 1X TaqMan Universal PCR Master Mix (Applied Biosystems), 450 nM of each primer, and 125 nM of each probe. The total reaction volume was 5μL/well. The reaction was initiated at 50°C for 2 min, followed by 95°C for 10 min and 50 cycles of 95°C for 15s and 60°C for 1 min. All real-time PCR experiments were carried out on an ABI PRISM 7900HT Sequence Detection System (Applied Biosystems), and the fluorescence data were collected by the "Absolute Quantification" application of the SDS 2.2.2 software (Applied Biosystems). The default baselines and manual threshold values were used. The number of wells which were positive for either chromosome 21 or chromosome 1 was recorded and subjected to SPRT analysis. One or more 384-well plates would be analyzed until disease classification was possible by SPRT.

IV. USING MICROFLUIDICS-BASED DIGITAL PCR

A. Digital RNA-SNP

[0187] This example demonstrates the performance of digital PCR analysis using microfluidics-based digital PCR. A variant of this approach is illustrated here, by way of example but not by way of limitation to, using a Fluidigm BioMark™ System. This system can perform over 9000 digital PCRs per run.

[0188] Placental tissues and maternal peripheral blood samples were obtained from pregnant women carrying euploid or trisomy 21 fetuses. Genotyping of the rs8130833 SNP on the *PLAC4* gene was carried out in placental DNA samples by primer extension followed by mass spectrometry. RNA was extracted from the placental and maternal plasma samples.

[0189] All RNA samples were reverse transcribed with a gene-specific reverse transcription primer (5′-AGTATATAGAACCATGTTTAGGCCAGA-3′) using the ThermoScript reverse transcriptase (Invitrogen). For the placental cDNA samples, serial dilutions were performed such that the subsequent digital PCR amplification could be performed at approximately one template molecule per well.

[0190] Digital PCR was conducted on the BioMark System™ (Fluidigm) with a 12.765 Digital Array (Fluidigm). Each Digital Array consists of 12 panels for accommodating 12 sample-assay mixtures. Each panel is further partitioned into 765 wells for carrying out a 7-nL reaction/well. The rs8130833 SNP region on the *PLAC4* gene was amplified by the forward primer (5′-TTTGTATTGCAACACCATTTGG-3′) and the gene-specific reverse transcription primer described above. Two allele-specific TaqMan probes targeting each of the two alleles of the rs8130833 SNP were designed. Their sequences were 5′-(FAM)TCGTCGTCTAACTTG(MGBNFQ)-3′ and 5′-(VIC)ATTCGTCATCTAACTTG(MGBNFQ) for the G and A alleles, respectively. The reaction for one array panel was set up using 2X TaqMan Universal PCR Master Mix in a reaction volume of 10 μL. Each reaction contains 1X TaqMan Universal PCR Master Mix, 572 nM of each primer, 53.5 nM of the allele-G-specific probe, 178.5 nM of the allele-A-specific probe and 3.5 μL of the cDNA sample. One reaction panel was used for each placental cDNA sample while 12 panels were used for each maternal plasma sample. The sample-assay mixtures were loaded into the Digital Array by a NanoFlex™ IFC controller (Fluidigm). The reaction was carried out in the BioMark™ System. The reaction was

initiated at 50°C for 2 min, followed by 95°C for 10 min and 40 cycles of 95°C for 15s and 57°C for 1 min.

[0191] Placental RNA samples from one euploid and two T21 heterozygous placentas were analyzed in a 765-well reaction panel. For each sample, the number of informative wells, comprising the ones positive for either the A or the G allele (but not both), was counted. The proportion of the overrepresented allele among all the informative wells (Pr) was determined. SPRT curves appropriate for the exact average reference template concentration per well (m_r) of these runs were applied to determine if the experimentally-obtained Pr indicated a euploid or T21 sample. As shown in FIG. 15A, all RNA samples were correctly classified using this approach.

[0192] We further tested the plasma RNA samples from four women carrying euploid fetuses and one woman carrying a trisomy 21 fetus. Each sample was analyzed in twelve 765-well reaction panels, i.e. 9180 reactions per plasma RNA sample. FIG. 15B shows the number of informative wells for each of the 12 panels for this plasma RNA sample. As shown in the table, the template concentration in the plasma sample was so diluted that the number of informative wells in any one reaction panel was not sufficient for the SPRT classification. The informative wells from three reaction panels had to be combined before this sample was classified as a euploid sample (FIG. 15C). FIG. 15C shows that using the aggregated data from two to twelve panels, all of the plasma cases could be correctly classified.

[0193] Compared with the manual method for performing digital PCR, this microfluidics-based method is much more rapid and less labor-intensive. The whole process could be completed in two and a half hours.

B. Digital RNA-SNP analysis for the prenatal detection of trisomy 18

[0194] In this example, we used a digital PCR-based allelic discrimination assay on *serpin peptidase inhibitor clade B (ovalbumin) member 2* (*SERPINB2*) mRNA, a placenta-expressed transcript on chromosome 18, to detect an imbalance in the ratio of polymorphic alleles for trisomy 18 fetuses. Extraction of DNA and RNA from placental tissue samples was performed using the QIAamp DNA Mini Kit (Qiagen, Hilden, Germany) and the TRIzol reagent (Invitrogen, Carlsbad, CA), respectively, as described in the manufacturers' protocols. The extracted placental RNA samples were subjected to DNase I treatment (Invitrogen) for removal of contaminating genomic DNA. Genotyping of the rs6098 SNP on

the SERPINB2 gene was carried out in placental tissue DNA samples with a Homogenous MassEXTEND (hME) assay using the MassARRAY Compact (Sequenom, San Diego) as previously described.

[0195] Reverse transcription for the SERPINB2 transcript was performed on the placental tissue RNA samples with a gene-specific primer 5′-CGCAGACTTCTCACCAAACA-3′ (Integrated DNA Technologies, Coralville, IA) using the ThermoScript reverse transcriptase (Invitrogen). All cDNA samples were diluted to a concentration such that the subsequent digital PCR amplification could be performed at an average concentration of one template molecule per reaction well. Digital PCR was set up using the TaqMan Universal PCR Master Mix (Applied Biosystems, Foster City, CA) and the BiomarkTM PCR Reagents (Fluidigm, San Francisco). The forward primer 5′- CTCAGCTCTGCAATCAATGC-3′ (Integrated DNA Technologies) and the reverse primer (identical to the gene-specific primer used for reverse transcription) were used at a concentration of 600 nM. The two TaqMan probes targeting the A or G allele of the rs6098 SNP on the SERPINB2 sequence were 5′-(FAM)CCACAGGGAATTATTT (MGBNFQ)-3′ and 5′-(FAM)CCACAGGGGATTATTT(MGBNFQ)-3′ (Applied Biosystems). FAM is 6-carboxyfluorescein and MGBNFQ is a minor groove-binding nonfluorescent quencher, and were used at concentrations of 300 nM and 500 nM, respectively. Each sample-reagent mix was partitioned into 765 reaction wells on a BiomarkTM 12.765 Digital Array using the NanoflexTM IFC Controller (Fluidigm). After partitioning, the array was placed in the BiomarkTM Real-time PCR System (Fluidigm) for thermal amplification and fluorescence detection. The reaction was initiated at 50oC for 2 min and continued at 95oC for 5 min followed by 45 cycles of 95oC for 15 sec and 59 oC for 1 min. After amplification, the number of informative wells (one that was positive for either the A or G allele only) and the number of wells positive for both alleles were counted and subjected to sequential probability ratio test (SPRT) analysis.

[0196] For a heterozygous euploid fetus, the A and G alleles should be equally represented (1:1) in the fetal genome, whereas for trisomy 18, there would be an additional copy of one allele thus giving a ratio of 2:1 in the fetal genome. A series of SPRT curves were generated for interpretation of different samples. These curves illustrate the expected proportion of informative wells positive for the overrepresented allele P_r (y-axis) for a given total number of informative wells (x-axis) needed for classification. For each sample, the experimentally derived P_r was compared with the expected P_r value. Samples above the upper curve were

classified as trisomy 18, whereas those below the bottom curve were classified as euploid. The area between the two curves is the unclassifiable region.

[0197] The feasibility of digital RNA-SNP analysis for the detection of fetal trisomy 18 was demonstrated by using the rs6098 SNP on the SERPINB2 gene. Placental tissue DNA samples from subjects with euploid and trisomy 18 fetuses were first genotyped by mass spectrometry for identifying heterozygous cases. Nine euploid and three trisomy 18 heterozygous placentas were found and subjected to digital RNA-SNP analysis. For each sample, P_r and m_r were calculated, and the SPRT curve for this m_r value was used for disease classification. As shown in FIG. 16A, all samples were correctly classified. The P_r values for trisomy 18 placentas were above the unclassifiable region, whereas those for euploid placentas fell below this region.

[0198] Samples with SPRT curves based on $m_r = 0.1$, 0.2, and 0.3 are illustrated in FIG. 16B. These data suggest that the digital RNA-SNP method is a valuable diagnostic tool for trisomy 18 pregnancies. The two curves represent the boundaries for the unclassifiable region. Samples with data points above the upper curve were classified as aneuploid, whereas those with data points below the bottom curve were classified as euploid.

C. Digital RCD analysis

[0199] This example demonstrates the performance of digital RCD analysis using microfluidics-based digital PCR. A variant of this approach is illustrated here, by way of example but not by way of limitation to, using a Fluidigm BioMark™ System. This system can perform over 9000 digital PCRs per run.

[0200] Placental tissues, maternal blood cell and amniotic fluid samples were obtained from pregnant women carrying euploid or trisomy 21 (T21) fetuses. Placental DNA from 10 T21 and 10 euploid cases was each mixed with an equal amount of euploid maternal blood cell DNA, thus producing twenty 50% DNA mixtures. To ensure accurate fetal proportion in the mixture samples, the extracted DNA was first quantified by optical density (OD) measurement at 260 nm. They were then digitally quantified by the BioMark™ System (Fluidigm) using the 12.765 Digital Arrays (Fluidigm). The assay for quantifying the samples was the same as described below except that only the probe for the reference chromosome was used.

[0201] The chromosome dosage in the 50% DNA mixtures and amniotic fluid samples was determined by digital PCR analysis of a nonpolymorphic chromosome 21 locus relative to one located on chromosome 1. A 101-bp amplicon of a pair of paralogous loci on chromosome 21 and 1 was first co-amplified by forward primer 5'-GTTGTTCTGCAAAAAACCTTCGA-3' and reverse primer 5'-CTTGGCCAGAAATACTTCATTACCATAT-3'. Two chromosome-specific TaqMan probes were designed to distinguish between the chromosome 21 and 1 paralogs, and their sequences were 5'-(FAM)TACCTCCATAATGAGTAAA(MGBNFQ)-3' and 5'-(VIC)CGTACCTCTGTAATGTGTAA(MGBNFQ)-3', respectively. The use of paralogous loci was used here by way of example only. In other words, non-paralogous loci could also be used for such analysis.

[0202] In order to demonstrate the use of the digital RCD approach to detect trisomy 18 (T18), another assay targeting paralogous sequences on chromosome 21 and 18 was designed. A 128-bp amplicon of the paralogous loci on chromosome 21 and 18 was first co-amplified by forward primer 5'-GTACAGAAACCACAAACTGATCGG-3' and reverse primer 5'-GTCCAGGCTGTGGGCCT-3'. Two chromosome-specific TaqMan probes were designed to distinguish between the chromosome 21 and 18 paralogs, and their sequences were 5'-(FAM)AAGAGGCGAGGCAA(MGBNFQ)-3' and 5'-(VIC)AAGAGGACAGGCAAC(MGBNFQ)-3', respectively. The use of paralogous loci was used here by way of example only. In other words, non-paralogous loci could also be used for such analysis.

[0203] All experiments were carried out on the BioMark™ System (Fluidigm) using the 12.765 Digital Arrays (Fluidigm). The reaction for one panel was set up using 2X TaqMan Universal PCR Master Mix (Applied Biosystems) in a reaction volume of 10 μL. Each reaction contained 1X TaqMan Universal PCR Master Mix, 900 nM of each primer, 125 nM of each probe and 3.5 μL of a 50% placental/maternal blood cell DNA sample. The sample/assay mixture was loaded into the Digital Array by the NanoFlex™ IFC controller (Fluidigm). The reaction was carried out on the BioMark™ System for detection. The reaction was initiated at 50°C for 2 min, followed by 95°C for 10 min and 40 cycles of 95°C for 15s and 57°C for 1 min.

[0204] The euploid and T21 50% placental/maternal blood cell DNA samples were analyzed on the digital arrays with the chr21/chr1 assay. For each sample, the number of

informative wells, comprising the ones positive for only chr21 or chr1 markers (but not both), was counted. The proportion of the overrepresented marker among all the informative wells (P_r) was determined. SPRT curves appropriate for the exact average reference template concentration per well (m_r) for any one of the digital PCR panels were applied to determine if the experimentally-obtained P_r indicated a euploid or T21 sample. Data were aggregated from extra panels for samples which remained unclassified until a decision could be made. As shown in FIG. 17, all 50% placental/maternal blood cell DNA samples were correctly classified using this approach with data ranging from one to four panels needed. A SPRT curve was also plotted to show the decision boundaries for correct classification, as shown in FIG. 18.

[0205] We further applied the RCD analysis on amniotic fluid samples obtained from 23 women carrying euploid fetuses and 6 women carrying T21 fetuses. Each sample was analyzed in a single 765-well reaction panel with the chr21/chr1 assay. FIG. 19 shows the SPRT classification summary. As shown in FIG. 19, all the 29 samples were classified correctly. The digital RCD method is thus an alternative approach for the detection of trisomies using microsatellite (Levett LJ, et al. A large-scale evaluation of amnio-PCR for the rapid prenatal diagnosis of fetal trisomy. Ultrasound Obstet Gynecol 2001; 17: 115-8) or single nucleotide polymorphism (SNP) (Tsui NB, et al. Detection of trisomy 21 by quantitative mass spectrometric analysis of single-nucleotide polymorphisms. Clin Chem 2005; 51: 2358-62) markers or real-time non-digital PCR (Zimmermann B, et al. Novel real-time quantitative PCR test for trisomy 21. Clin Chem 2002; 48: 362-3) in miscellaneous sample types used for prenatal diagnosis, such as amniotic fluid and chorionic villus biopsies.

[0206] In an attempt to detect T18 cases, we applied the chr21/chr18 assay on 3 euploid and 5 T18 placental DNA samples. The proportion of the overrepresented marker among all the informative wells (P_r) was calculated. All of them were classified correctly except one T18 case was misclassified as euploid. The results were summarized in FIG. 20.

V. USING MULTIPLEX DIGITAL RCD ASSAYS ON MASS-SPECTROMETRIC PLATFORM

[0207] The number of plasma DNA molecules that are present per unit volume of maternal plasma is limited (Lo YMD. *et al.* 1998 *Am J Hum Genet* 62, 768-7758). For example, in early pregnancy, the median maternal plasma concentration of an autosomal locus, the *β-globin* gene, has been shown to be 986 copies/mL, with contributions from both the fetus

and mother (Lo YMD. *et al.* 1998 *Am J Hum Genet* 62, 768-7758). To capture 7,680 molecules, DNA extracted from some 8 mL of maternal plasma would be needed. This volume of plasma, obtainable from ~15 mL of maternal blood, is at the limit of routine practice. However, we envision that multiple sets of chr21 and reference chromosome targets can be combined for digital RCD analysis. For 5 pairs of chr21 and reference chromosome targets, just 1.6 mL of maternal plasma would be needed to provide the number of template molecules needed for analysis. Multiplex single molecule PCR could be performed. The robustness of such multiplex single molecule analysis has been demonstrated previously for single molecule haplotyping (Ding, C. and Cantor, CR. 2003 *Proc Natl Acad Sci U S A* 100, 7449-7453).

[0208] In one example, placental tissues and maternal blood cell samples were obtained from pregnant women carrying euploid or trisomy 21 (T21) fetuses. 5 euploid and 5 T21 placental DNA samples were each mixed with equal proportions of maternal blood cell DNA to produce 10 DNA mixtures mimicking plasma samples with 50% fetal DNA. To ensure accurate fetal proportion in the mixture samples, the extracted DNA was first quantified by optical density (OD) measurement at 260 nm. They were then digitally quantified by real-time PCR in 384-well format. The assay for quantifying the samples was the same as described in the previous example of digital RCD analysis.

[0209] The chromosome dosage in the 50% mix was determined by digital PCR analysis of a nonpolymorphic chromosome 21 locus relative to one located on chromosome 1. The method is called Digital Relative Chromosome Dosage (RCD) analysis. A 121-bp amplicon (inclusive of a 10-mer on each primer) of a pair of paralogous loci on chromosome 21 and 1 was co-amplified by forward primer 5'-ACGTTGGATGGTTGTTCTGCAAAAAACCTTCGA-3' and reverse primer 5'-ACGTTGGATGCTTGGCCAGAAATACTTCATTACCATAT-3'. An extension primer which targets the base differences between chromosome 21 and chromosome 1 was designed, and its sequence is 5'-CTCATCCTCACTTCGTACCTC-3'.

[0210] In order to demonstrate the utility of multiplexing digital PCR assays to detect T21 cases, another digital RCD assay targeting paralogous sequences on chromosome 21 and 18 was designed. A 148-bp amplicon (inclusive of a 10-mer on each primer) of the paralogous loci on chromosome 21 and 18 was co-amplified by forward primer 5'-ACGTTGGATGGTACAGAAACCACAAACTGATCGG-3' and reverse primer 5'-

ACGTTGGATGGTCCAGGCTGTGGGCCT-3'. An extension primer which targets the base differences between chromosome 21 and chromosome 18 was designed, and its sequence is 5'-ACAAAAGGGGGAAGAGG-3'.

[0211] Multiplex digital RCD analysis was performed using primer extension protocol. PCR reaction was set up using GeneAmp PCR Core Reagent Kit (Applied Biosystems) in a reaction volume of 5 μL. Each reaction contained 1X Buffer II, 2 mM $MgCl_2$, 200 μM dNTP mix, 0.2 U AmpliTaq Gold, 200 nM of each of the 4 primers and the 50% DNA mix. The assay/sample mixture was dispensed into 384-well PCR plate and the reaction was initiated at 50°C for 2 min, followed by 95°C for 10 min and 40 cycles of 95°C for 15s and 57°C for 1 min.

[0212] PCR products were subjected to shrimp alkaline phosphatase (SAP) treatment to remove unincorporated dNTPs. The mixture was incubated at 37°C for 40 min followed by 85°C for 5 min. Primer extension reaction was then carried out. In brief, 771 nM of extension primer from chr21/chr1 assay, 1.54 μM of extension primer from chr21/chr18 assay, 0.67 U Thermosequenase (Sequenom), and 64 μM each of ddCTP, ddGTP, dATP and dTTP in an extension cocktail were added to the SAP-treated PCR products. The reaction conditions were 94°C for 2 min, followed by 94°C for 5 s, 50°C for 5 s, and 72°C for 5 s for 80 cycles. 16 μL of water and 3 mg of the Clean Resin (Sequenom) were added to the extension products for a final clean up. The mixtures were mixed in a rotator for 20 to 30 min, followed by a centrifugation step at 361 g for 5 min. Fifteen to 25 nL of the final products were dispensed onto a SpectroCHIP (Sequenom) by a MassARRAY Nanodispenser S (Sequenom). Data acquisition from the SpectroCHIP was done in the MassARRAY Analyzer Compact Mass Spectrometer (Sequenom). Mass data were imported into the MassARRAY Typer (Sequenom) software for analysis.

[0213] The five euploid and five T21 50% placental/maternal DNA samples were analyzed with the duplex RCD assay. For each sample, the number of informative wells from individual assay, comprising the ones positive for only chr21 or chr1 or chr18 markers, was counted. The proportion of the chr21 marker among all the informative wells (P_r) was calculated separately for each RCD assay. Sequential probability ratio test (SPRT) was then applied to determine if the P_r indicated a euploid or T21 sample. By doing so, the number of wells required was reduced as each plate was counted twice.

[0214] The chr21/chr1 assay was usually applied first. If the sample remained unclassified, then the values from the chr21/chr18 assay would be added for further calculations. Extra plates were used for samples which remained unclassified until a decision could be made. As shown in FIG. 21, all euploid 50% mix samples were correctly classified using a single 384-well plate. Several T21 cases required 2 or more plates for correct classification. If only one assay was used, a greater number of plates would be needed to attain the number of informative wells required when classification was achieved. For example, data for case N0230 was unclassifiable when either of the RCD assays was used alone. However, correct classification was achieved when data from the two assays were combined. If the duplex RCD assays were not used, additional plates of analyses would be needed. We would expect a further reduction of well number with a higher level of multiplexing of assays.

[0215] In another example, we developed a 4-plex assay targeting 4 different amplicons on chromosome 21 and their corresponding paralogous partners located on other non-chromosome-21 autosomes. This 4-plex assay was used in digital RCD analysis followed by SPRT classification of samples from euploid and trisomy 21 pregnancies. DNA extractions from placental samples were performed using the QIAamp tissue kit (Qiagen, Hilden, Germany).

[0216] All placental and maternal buffy coat DNA samples used in this study were first quantified by the NanoDrop spectrophotometer (NanoDrop Technology, Wilmington, DE). The DNA concentration was converted to genome equivalent (GE)/μL using a conversion of 6.6 pg/cell. The amount of DNA corresponding to approximately one template per well was determined by serially diluting the DNA samples. Under such a condition, we would expect approximately 37% of the wells to show negative amplification. In multiplex digital RCD analysis, 4 sets of paralogous sequence targets were selected: the paralogous loci on chromosome 21 and 1 were co-amplified by the forward primer 5'-ACGTTGGATGTTGATGAAGTCTCATCTCTACTTCG -3' and the reverse primer 5'-ACGTTGGATGCAATAAGCTTGGCCAGAAATACT -3', resulting in an amplicon of 81 bp. The paralogous loci on chromosome 21 and 7 were co-amplified by the forward primer 5'- ACGTTGGATGGAATTTAAGCTAAATCAGCCTGAACTG-3' and the reverse primer 5'- ACGTTGGATGGTTTCTCATAGTTCATCGTAGGCTTAT -3', resulting in an amplicon of 82 bp. The paralogous loci on chromosome 21 and 2 were co-amplified by the forward primer 5'- ACGTTGGATGTCAGGCAGGGTTCTATGCAG -3' and the reverse primer 5'-ACGTTGGATGAGGCGGCTTCCTGGCTCTTG -3', resulting in an amplicon of 101 bp.

The paralogous loci on chromosome 21 and 6 were co-amplified by the forward primer 5'-ACGTTGGATGGCTCGTCTCAGGCTCGTAGTT -3' and the reverse primer 5'-ACGTTGGATGTTTCTTCGAGCCCTTCTTGG -3', resulting in an amplicon of 102 bp. Each reaction contained 10X buffer II (Applied Biosystems), $MgCl_2$ and 100 nM of each primer. The total reaction volume was 5 μL/well. The reaction was initiated at 95°C for 5 min, followed by 45 cycles of 95°C for 30 sec, 62 °C for 30 sec and 72°C for 30 sec, and a final extension at 72 °C for 7 min. All conventional PCR amplifications were carried out on a GeneAmp PCR System 9700 (Applied Biosystems). The unincorporated nucleotides were deactivated by shrimp alkaline phosphatase (SAP) treatment. Each reaction contained 10X SAP buffer (Sequenom) and SAP enzyme (Sequenom). 2 μL of SAP mix was added to each PCR. The SAP reaction was incubated at 37 °C for 40 min and 85 °C for 5 min. After the SAP treatment, primer extension reaction was carried out on the PCR products using the iPLEX Gold kit (Sequenom). . The paralogous sequence mismatches (PSMs) on the paralogous loci on chromosomes 21 and 1 were interrogated by the extension primer 5'-GTCTCATCTCTACTTCGTACCTC -3'. The PSMs on the paralogous loci on chromosomes 21 and 7 were interrogated by the extension primer 5'-TTTTACGCTGTCCCCATTT-3'. The PSMs on the paralogous loci on chromosomes 21 and 2 were interrogated by the extension primer 5'-GGTCTATGCAGGAGCCGAC-3'. The PSMs on the paralogous loci on chromosomes 21 and 6 were interrogated by the extension primer 5'-TGGGCGCGGGAGCGGACTTCGCTGG-3'. Each reaction contained 10X iPLEX buffer (Sequenom), iPLEX termination mix (Sequenom), iPLEX enzyme (Sequenom) and 343 nM of each extension primer, except for the extension primer for PSMs on chromosomes 21 and 6 which was used at 1.03 μM. 2 μL of iPLEX mix was added to 5 μL of PCR product. The iPLEX reaction was cycled according to a 200-short-cycle program. Briefly, the samples were first denatured at 94 °C for 35 sec, followed by annealing at 52° C for 5 sec and extension at 80°C for 5 sec. The annealing and extension cycle was repeated four more times for a total of five cycles and then looped back to a 94° C denaturing step for 5 sec, after which was the 5-cycle annealing and extension loop again. The five annealing and extension cycles with the single denaturing step were repeated 39 times for a total of 40. A final extension at 72° C for 3 min was performed. The iPLEX reaction products were diluted with 16 μL water and desalted by 6 mg resin for each PCR. The 384-well plate was centrifuged at 1600 *g* for 3 min before dispensing onto the SpectroCHIP (Sequenom) and the matrix-assisted laser

desorption/ionization time-of-flight (MALDI-TOF) mass spectrometry MS Analysis (Sequenom).

[0217] The number of wells which were positive for only chromosome 21 or only the reference chromosome for each of the four assays was independently recorded. For each assay, the Poisson corrected numbers of molecules for chromosome 21 and the reference chromosome were calculated. The sum of the Poisson corrected number of molecules for chromosome 21 as well as the sum of the Poisson corrected number of reference chromosomes from all four assays were calculated and deemed as the informative counts for the 4-plex assay. The P_r value was the chromosome 21 count for the 4-plex assay divided by the sum of the chromosome 21 and reference chromosome counts for the 4-plex assay. The experimentally derived P_r values were subjected to SPRT analysis. One or more 384-well plates would be analyzed until disease classification was possible by SPRT. A total of two 50% euploid placental genomic DNA/50% maternal buffy coat DNA mix and two 50% trisomy 21 placental genomic DNA/50% maternal buffy coat DNA mix were analyzed.

[0218] The experimentally derived P_r value would be compared with the expected value of P_r to test the null or alternative hypotheses. Alternatively, neither the null or alternative hypothesis could be accepted if the P_r for the given number of informative counts has not yet reached the required level of statistical confidence for disease classification. These cases were deemed unclassifiable until more data were available.

[0219] The results and the SPRT classification of each sample are tabulated in FIGS. 22A and 22B. The two euploid samples required 2 and 5 384-well multiplex digital RCD analyses before SPRT classification could be reached. Data from none of the individual member of the 4-plex assay allowed disease classificaiton by SPRT. Both trisomy 21 samples were each correctly classified with just one 384-well multiplex digital RCD analysis. Similarly, data from none of the individual member of the 4-plex assay allowed disease classificaiton by SPRT. However, the composite counts from the 4-plex assay allowed correct SPRT classification. These data illustrated that by using multiplex digital RCD, the effective number of informative counts were substantially increased for a given number of digital PCR analyses performed as compared to the use of a single-plex digital RCD assay.

VI. USING DIGITAL EPIGENETIC RELATIVE CHROMOSOME DOSAGE

[0220] Here we outline an approach called digital epigenetic relative chromosome dosage (digital ERCD) in which epigenetic markers exhibiting a fetal-specific DNA methylation pattern, or other epigenetic changes, on a chromosome involved in a chromosomal aneuploidy (e.g. chromosome 21 in trisomy 21) and on a reference chromosome, are subjected to digital PCR analysis. The ratio of the number of wells positive for the chromosome 21 epigenetic marker to that positive for the reference chromosome epigenetic marker in plasma DNA extracted from pregnant women bearing normal fetuses will give us the reference range. The ratio will be expected to be increased if the fetus has trisomy 21. It is obvious to those of skill in the art that more than one chromosome 21 markers and more than one reference chromosome markers could be used in this analysis.

[0221] One example of a gene on chromosome 21 which exhibits a fetus (placenta)-specific methylation pattern is the *Holocarboxylase synthetase (HLCS)* gene. *HLCS* is hypermethylated in the placenta, but hypomethylated in maternal blood cells; and is covered in U.S. patent application 11/784499, which is incorporated herein by reference. One example of a gene on a reference chromosome which exhibits a fetus (placenta)-specific methylation pattern is the *RASSF1A* gene on chromosome 3 [10]. *RASSF1A* is hypermethylated in the placenta but is hypomethylated in maternal blood cells, see U.S. patent application 11/784501, which is incorporated herein by reference.

[0222] In the application of hypermethylated *HLCS* and hypermethylated *RASSF1A* to digital PCR detection of trisomy 21 in a fetus using maternal plasma, maternal peripheral blood is first collected. Then the blood is subjected to centrifugation and the plasma is harvested. DNA from the plasma is then extracted using techniques well-known to those of skill in the art, such as using a QIAamp Blood kit (Qiagen). The plasma DNA is then subjected to digestion using one or more methylation-sensitive restriction enzymes, such as *Hpa*II and *Bst*UI. These methylation-sensitive restriction enzyme(s) will cut the maternal, nonmethylated versions of these genes, while leaving the fetal hypermethylated sequences intact. The digested plasma DNA sample is then diluted to an extent that on average approximately 0.2 to 1 molecule of either the restriction enzyme treated but intact *HLCS* or *RASSF1A* sequences will be detected per reaction well. Two real-time PCR systems will be used to amplify the diluted DNA, one with two primers and one TaqMan probe specific to the *HLCS* gene, encompassing the region that will be cut by the restriction enzyme(s) if the

sequence in unmethylated; and the other one towards the *RASSF1A* gene, similarly with two primers and one TaqMan probe. With regard to the latter *RASSF1A* primer/probe set, one example has been described by Chan et al 2006, Clin Chem 52, 2211-2218. The TaqMan probes towards the *HLCS* and *RASSF1A* targets will have different fluorescent reporters, such as FAM and VIC, respectively. A 384-well plate is then used to perform the digital PCR experiment. The number of wells scored positive for just *HLCS* and those scored positive for just *RASSF1A* will be counted, and a ratio of these counts will be taken. The *HLCS*:*RASSF1A* ratio will be expected to be higher for maternal plasma taken from a pregnant woman carrying a trisomy 21 fetus, when compared with one carrying a normal euploid fetus. The degree of overrepresentation will be dependent on the average reference template concentration per well in the digital PCR run.

[0223] Other methods for scoring these results will be possible, for the example the counting of the number of wells positive for *HLCS*, irrespective of the concurrent positivity for *RASSF1A*; and vice versa for *RASSF1A*, irrespective of the concurrent positivity for *HLCS*. Furthermore, in replacement of calculating the ratio, either the total number, or the difference in the *HLCS* and *RASSF1A* counts could be used to indicate the trisomy 21 status of a fetus.

[0224] Apart from doing the digital PCR in plates, it will also be obvious to those of skill in the art that other variants of digital PCR can be used, e.g. microfluidics chips, nanoliter PCR microplate systems, emulsion PCR, polony PCR and rolling-circle amplification, primer extension and mass spectrometry, etc. These variants of digital PCR are named by way of examples, and not as limitations.

[0225] Apart from real-time PCR, it will also be obvious to those of skill in the art that methods such as mass spectrometry can be used to score the digital PCR results.

[0226] Apart from using methylation-sensitive restriction enzymes to differentiate the fetal and maternal versions of *HLCS* and *RASSF1A*, it will be obvious to those of skill in the art that other methods for ascertaining the methylation status would also be applicable, e.g. bisulfite modification, methylation-specific PCR, immunoprecipitation using antibody to methylated cytosine, mass spectrometry, etc.

[0227] It will also be obvious to those of skill in the art that the approach illustrated in this example and other examples in this patent application can be used in the other bodily fluids in

which fetal DNA may be found, including maternal urine, amniotic fluid, transcervical washings, chorionic villus, maternal saliva, etc.

VII. MASSIVELY PARALLEL GENOMIC SEQUENCING USING EMULSION PCR AND OTHER STRATEGIES

[0228] Here we shall describe another example whereby a digital readout of nucleic acid molecules can be used for the detection of fetal chromosomal aneuploidies, e.g. trisomy 21, in maternal plasma. Fetal chromosomal aneuploidy results from abnormal dose(s) of a chromosome or chromosomal region. It is desirable that noninvasive tests have high sensitivity and specificity to minimize false diagnoses. However, fetal DNA is present in low absolute concentration and represent a minor portion of all DNA sequences in maternal plasma and serum. Hence, the number of digital PCR sampling targeting specific gene loci cannot be increased infinitely within the same biological specimen. Hence, the analysis of multiple sets of specific target loci may be used to increase the amount of data that could be obtained from a specimen without increasing the number of digital PCR sampling performed.

[0229] Accordingly, embodiments allow the noninvasive detection of fetal chromosomal aneuploidy by maximizing the amount of genetic information that could be inferred from the limited amount of fetal nucleic acids which exist as a minor population in a biological sample containing maternal background nucleic acids. In one aspect, the amount of genetic information obtained is sufficient for accurate diagnosis yet not overly excessive so as to contain costs and the amount of input biological sample required.

[0230] Massively parallel sequencing, such as that achievable on the 454 platform (Roche) (Margulies, M. *et al.* 2005 *Nature* 437, 376-380), Illumina Genome Analyzer (or Solexa platform) or SOLiD System (Applied Biosystems) or the Helicos True Single Molecule DNA sequencing technology (Harris TD et al. 2008 Science, 320, 106-109), the single molecule, real-time (SMRT™) technology of Pacific Biosciences, and nanopore sequencing (Soni GV and Meller A. 2007 Clin Chem 53: 1996-2001), allow the sequencing of many nucleic acid molecules isolated from a specimen at high orders of multiplexing in a parallel fashion (Dear Brief Funct Genomic Proteomic 2003; 1: 397-416). Each of these platforms sequences clonally expanded or even non-amplified single molecules of nucleic acid fragments.

[0231] As a high number of sequencing reads, in the order of hundred thousands to millions or even possibly hundreds of millions or billions, are generated from each sample in each run,

the resultant sequenced reads form a representative profile of the mix of nucleic acid species in the original specimen. For example, the haplotype, trascriptome and methylation profiles of the sequenced reads resemble those of the original specimen (Brenner et al Nat Biotech 2000; 18: 630-634; Taylor et al Cancer Res 2007; 67: 8511-8518). Due to the large sampling of sequences from each specimen, the number of identical sequences, such as that generated from the sequencing of a nucleic acid pool at several folds of coverage or high redundancy, is also a good quantitative representation of the count of a particular nucleic acid species or locus in the original sample.

[0232] In one embodiment, random sequencing is performed on DNA fragments that are present in the plasma of a pregnant woman, and one obtains genomic sequences which would originally have come from either the fetus or the mother. Random sequencing involves sampling (sequencing) a random portion of the nucleic acid molecules present in the biological sample. As the sequencing is random, a different subset (fraction) of the nucleic acid molecules (and thus the genome) may be sequenced in each analysis. Embodiments will work even when this subset varies from sample to sample and from analysis to analysis, which may occur even using the same sample. Examples of the fraction are about 0.1%, 0.5%, or 1% of the genome. In other embodiments, the fraction is at least any one of these values.

[0233] A bioinformatics procedure may then be used to locate each of these DNA sequences to the human genome. It is possible that a proportion of such sequences will be discarded from subsequent analysis because they are present in the repeat regions of the human genome, or in regions subjected to inter-individual variations, e.g. copy number variations. An amount of the chromosome of interest and of one or more other chromosomes may thus be determined.

[0234] In one embodiment, a parameter (e.g. a fractional representation) of a chromosome potentially involved in a chromosomal aneuploidy, e.g. chromosome 21 or chromosome 18 or chromosome 13, may then be calculated from the results of the bioinformatics procedure. The fractional representation may be obtained based on an amount of all of the sequences (e.g. some measure of all of the chromosomes) or a particular subset of chromosomes (e.g. just one other chromosome than the one being tested.)

[0235] This fractional representation is then compared to a reference range established in pregnancies involving normal (i.e. euploid) fetuses. It is possible that in some variants of the

procedure, the reference range would be adjusted in accordance with the fractional concentration of fetal DNA (f) in a particular maternal plasma sample. The value of f can be determined from the sequencing dataset, e.g. using sequences mappable to the Y chromosome if the fetus is male. The value of f may also be determined in a separate analysis, e.g. using fetal epigenetic markers (Chan KCA et al 2006 Clin Chem 52, 2211-8) or from the analysis of single nucleotide polymorphisms.

[0236] In one aspect, even when a pool of nucleic acids in a specimen is sequenced at <100% genomic coverage, and among the proportion of captured nucleic acid molecules, most of each nucleic acid species is only sequenced once, dosage imbalance of a particular gene locus or chromosome can also be quantitatively determined. In other words, the dosage imbalance of the gene locus or chromosome is inferred from the percentage representation of the said locus among all mappable sequenced tags of the specimen.

[0237] In one aspect for the massively parallel genomic sequencing approach, representative data from all of the chromosomes may be generated at the same time. The origin of a particular fragment is not selected ahead of time. The sequencing is done at random and then a database search may be performed to see where a particular fragment is coming from. This is contrasted from situations when a specific fragment from chromosome 21 and another one from chromosome 1 are amplified.

[0238] In one example, a proportion of such sequences would be from the chromosome involved in an aneuploidy such as chromosome 21 in this illustrative example. Yet other sequences from such a sequencing exercise would be derived from the other chromosomes. By taking into account of the relative size of chromosome 21 compared with the other chromosomes, one could obtain a normalized frequency, within a reference range, of chromosome 21-specific sequences from such a sequencing exercise. If the fetus has trisomy 21, then the normalized frequency of chromosome 21-derived sequences from such a sequencing exercise will increase, thus allowing the detection of trisomy 21. The degree of change in the normalized frequency will be dependent on the fractional concentration of fetal nucleic acids in the analyzed sample.

[0239] In one embodiment, we used the Illumina Genome Analyzer for single-end sequencing of human genomic DNA and human plasma DNA samples. The Illumina Genome Analyzer sequences clonally-expanded single DNA molecules captured on a solid surface termed a flow cell. Each flow cell has 8 lanes for the sequencing of 8 individual

specimens or pools of specimens. Each lane is capable of generating ~ 200Mb of sequence which is only a fraction of the 3 billion basepairs of sequences in the human genome. Each genomic DNA or plasma DNA sample was sequenced using one lane of a flow cell. The short sequence tags generated were aligned to the human reference genome and the chromosomal origin was noted. The total number of individual sequenced tags aligned to each chromosome were tabulated and compared with the relative size of each chromosome as expected from the reference human genome or non-disease representative specimens. Chromosome gains or losses were then identified.

[0240] The described approach is only one exemplification of the presently described gene/chromosome dosage strategy. Alternatively, paired end sequencing could be performed. Instead of comparing the length of the sequenced fragments from that expected in the reference genome as described by Campbell et al (Nat Genet 2008; 40: 722-729), the number of aligned sequenced tags were counted and sorted according to chromosomal location. Gains or losses of chromosomal regions or whole chromosomes were determined by comparing the tag counts with the expected chromosome size in the reference genome or that of a non-disease representative specimen

[0241] In another embodiment, the fraction of the nucleic acid pool that is sequenced in a run is further sub-selected prior to sequencing. For example, hybridization based techniques such as oligonucleotide array could be used to first sub-select for nucleic acid sequences from certain chromosomes, e.g. a potentially aneuploid chromosome and other chromosome(s) not involved in the aneuploidy tested. Another example is that a certain sub-population of nucleic acid sequences from the sample pool is sub-selected or enriched prior to sequencing. For example, it has been reported that fetal DNA molecules in maternal plasma are comprised of shorter fragments than the maternal background DNA molecules (Chan et al Clin Chem 2004; 50: 88-92). Thus, one may use one or more methods known to those of skill in the art to fractionate the nucleic acid sequences in the sample according to molecule size, e.g. by gel electrophoresis or size exclusion columns or by microfluidics-based approach. Yet, alternatively, in the example of analyzing cell-free fetal DNA in maternal plasma, the fetal nucleic acid portion could be enriched by a method that suppresses the maternal background, such as by the addition of formaldehyde (Dhallan et al JAMA 2004; 291: 1114-9).

[0242] Other single molecule sequencing strategies such as that by the Roche 454 platform, the Applied Biosystems SOLiD platform, the the Helicos True Single Molecule DNA

sequencing technology, the single molecule, real-time (SMRT™) technology of Pacific Biosciences, and nanopore sequencing could similarly be used in this application.

[0243] Examples of results and a further discussion (e.g. for sequencing and calculating parameters) may be found in concurrently filed application "DIAGNOSING FETAL CHROMOSOMAL ANEUPLOIDY USING GENOMIC SEQUENCING," (Attorney Docket No. 016285-005220US), which is incorporated by reference. Note that methods described herein for determining a cutoff value may be implemented when the reaction is sequencing, e.g. as described in this section.

[0244] The determination of the fractional concentration of fetal DNA in maternal plasma can also be done separate to the sequencing run. For example, the Y chromosome DNA concentration could be pre-determined using real-time PCR, microfluidics PCR or mass spectrometry. In fact, fetal DNA concentration could be determined using loci other than the Y chromosome and applicable to female fetuses. For example, Chan et al showed that fetal-derived methylated *RASSF1A* sequences would be detected in the plasma of pregnant women in the background of maternally derived unmethylated *RASSF1A* sequences (Chan et al, Clin Chem 2006;52:2211-8). The fractional fetal DNA concentration can thus be determined by dividing the amount of methylated *RASSF1A* sequences by the amount of total *RASSF1A* (methylated and unmethylated) sequences.

[0245] It is expected that maternal plasma would be preferred over maternal serum for practicing our invention because DNA is released from the maternal blood cells during blood clotting. Thus, if serum is used, it is expected that the fractional concentration of fetal DNA will be lower in maternal plasma than maternal serum. In other words, if maternal serum is used, it is expected that more sequences would need to be generated for fetal chromosomal aneuploidy to be diagnosed, when compared with a plasma sample obtained from the same pregnant woman at the same time.

[0246] Yet another alternative way of determining the fractional concentration of fetal DNA would be through the quantification of polymorphic differences between the pregnant women and the fetus (Dhallan R, et al. 2007 *Lancet*, 369, 474-481). An example of this method would be to target polymorphic sites at which the pregnant woman is homozygous and the fetus is heterozygous. The amount of fetal-specific allele can be compared with the amount of the common allele to determine the fractional concentration of fetal DNA.

[0247] In contrast to the existing techniques for detecting chromosomal aberrations, including comparative genomic hybridization, microarray comparative genomic hybridization, quantitative real-time polymerase chain reaction, which detect and quantify one or more specific sequence(s), massively parallel sequencing is not dependent on the detection or analysis of predetermined or a predefined set of DNA sequences. A random representative fraction of DNA molecules from the specimen pool is sequenced. The number of different sequence tags aligned to various chromosomal regions is compared between specimens containing or not containing tumoral DNA. Chromosomal aberrations would be revealed by differences in the number (or percentage) of sequences aligned to any given chromosomal region in the specimens.

[0248] In another example the sequencing technique on plasma cell-free DNA may be used to detect the chromosomal aberrations in the plasma DNA for the detection of a specific cancer. Different cancers have a set of typical chromosomal aberrations. Changes (amplifications and deletions) in multiple chromosomal regions may be used. Thus, there would be an increased proportion of sequences aligned to the amplified regions and a decreased proportion of sequences aligned to decreased regions. The percentage representation per chromosome could be compared with the size for each corresponding chromosome in a reference genome expressed as percentage of genomic representation of any given chromosome in relation to the whole genome. Direct comparisons or comparisons to a reference chromosome may also be used.

VIII. MUTATION DETECTION

[0249] Fetal DNA in maternal plasma exists as a minor population, with an average of 3% to 6% of maternal plasma DNA being contributed by the fetus. Because of this reason, most of the previous work in the field has focused on the detection of DNA targets which the fetus has inherited from the father, and which are distinguishable from the majority maternal DNA background in maternal plasma. Examples of such previously detected targets include the *SRY* gene on the Y chromosome (Lo YMD et al. 1998 *Am J Hum Genet,* 62, 768-775) and the *RHD* gene when the mother is RhD-negative (Lo YMD et al. 1998 *N Engl J Med,* 339, 1734-1738.

[0250] For fetal mutation detection, previous strategies using maternal plasma are limited to autosomal dominant conditions in which the father is a carrier, the exclusion of autosomal recessive diseases by direct mutation detection when the father and mother carries different

mutations, or by linkage analysis (Ding C. et al 2004 *Proc Natl Acad Sci U S A* 101, 10762-10767). These previous strategies have significant limitations. For example, for a couple where both the male and female partners carry the same mutation, then it would be impossible to carry out meaningful prenatal diagnosis by direct mutation detection in maternal plasma.

[0251] Such a scenario is illustrated in FIG. 23. In this scenario, there will be three possible fetal genotypes, NN, NM and MM, where N represents the normal allele and M represents the mutant allele. Examples of mutant alleles include those causing cystic fibrosis, beta-thalassemia, alpha-thalassemia, sickle cell anemia, spinal muscular atrophy, congenital adrenal hyperplasia, etc. Other examples of such disorders can be found in the Online Mendelian Inheritance in Man (OMIM) www.ncbi.nlm.nih.gov/sites/entrez?db=OMIM&itool=toolbar. In maternal plasma, most of the DNA will be coming from the mother and would be NM. For any of the three fetal genotypes, there will not be any unique fetal allele which would allow its unique detection in maternal plasma. Thus, the conventional strategy cannot be applied here.

[0252] Embodiments described herein allow handing such scenarios. In the scenario where the mother and fetus are both NM, then the N allele and M allele will be in allelic balance. However, if the mother is NM and the fetus is NN, then there will be allelic imbalance in maternal plasma, with the N allele being overrepresented. On the other hand, if the mother is NM and the fetus is MM, then there will be allelic imbalance in maternal plasma, with the M allele being overrepresented. Thus, for fetal mutation detection, the null hypothesis refers to the absence of allelic imbalance when the fetus is of the NM genotype. The alternative hypothesis refers to the presence of allelic imbalance and the fetal genotype could be NN or MM depending on whether the N or M allele is overrepresented.

[0253] The presence or absence of allelic imbalance can be determined by digital PCR using embodiments described herein. In a first scenario, a particular volume of maternal plasma contains the DNA released from 100 cells, in which 50 are from the mother and 50 are from the fetus. Thus, the fractional concentration of fetal DNA in this volume of plasma is 50%. When the mother is of the genotype NM, then there will be 50 N alleles and 50 M alleles contributed by the mother. If the fetus is of the genotype NM, then there will be 50 N alleles and 50 M alleles contributed by the fetus. Therefore, there will be no allelic imbalance between the N allele and the M allele, a total of 100 copies each. On the other hand, if the

fetus is of the NN genotype, then there will be 100 fetal-derived N alleles in this volume of plasma. Thus, there will be a total of 150 N alleles to 50 M alleles. In other words, there will be allelic imbalance between N and M, with N being overrepresented at a ratio of 3:1 in relation to M.

[0254] In the converse situation, if the fetus is of the MM genotype, then there will be 100 fetal-derived M alleles in this volume of plasma. Thus, there will be 150 M alleles to 50 N alleles. In other words, there will be allelic imbalance between N and M, with M being overrepresented at a ratio of 3:1 in relation to N. Such allelic imbalance can be measured by digital PCR. The allele with the smaller number of positive wells is considered as the reference template. Similar to digital RNA-SNP and digital RCD analyses, the actual distribution of the alleles in the digital PCR experiment would be governed by the Poisson probability density function. Therefore, while the theoretical degree of allelic imbalance in the present scenario is 3:1, the expected degree of allelic imbalance would be dependent on the average template concentration per well during the digital PCR analysis. Thus interpretation cutoffs, such as for SPRT analysis, appropriate for the average reference template concentration per well (m_r) would need to be used for case classification.

[0255] Furthermore, the degree of allelic imbalance that needs to be measured is dependent on the fractional fetal DNA concentration. In contrast to the above example, let's consider a particular volume of maternal plasma contains the DNA released from 100 cells, in which 90 are from the mother and 10 are from the fetus. Thus, the fractional concentration of fetal DNA in this volume of plasma is 10%. When the mother is of the genotype NM, then there will be 90 N alleles and 90 M alleles contributed by the mother. If the fetus is of the genotype NM, then there will be 10 N alleles and 10 M alleles contributed by the fetus. Therefore, there will be no allelic imbalance between the N allele and the M allele, a total of 100 copies each. On the other hand, if the fetus is of the NN genotype, then there will be 20 fetal-derived N alleles in this volume of plasma. Thus, there will be a total of 110 N alleles to 90 M alleles.

[0256] In other words, there will be allelic imbalance between N and M, with N being overrepresented. In the converse situation, if the fetus is of the MM genotype, then there will be 20 fetal-derived M alleles in this volume of plasma. Thus, there will be 110 M alleles to 90 N alleles. In other words, there will be allelic imbalance between N and M, with M being overrepresented. The theoretical degree of allelic imbalance when the fetal DNA fractional concentration is 10% would be 110:90, which is different to the 3:1 ratio when there is 50%

fetal DNA as shown in the above example. Thus interpretation cutoffs, such as for SPRT analysis, appropriate for the fetal DNA fractional concentration would need to be used for case classification.

[0257] Thus, plasma DNA will be extracted. The amount of maternal and fetal DNA in the plasma sample will be quantified, for example by the real-time PCR assays previously established (Lo, *et al.* 1998 *Am J Hum Genet* 62, 768-775) or other types of quantifier well-known to those of skill in the art, e.g. SNP markers (Dhallan R et al. 2007 *Lancet,* 369, 474-481) and fetal epigenetic markers (Chan KCA et al. 2006 *Clin Chem,* 52, 2211-2218). The fetal DNA percentage will be calculated. Then the quantified plasma DNA sample is prepared (e.g. diluted or concentrated) such that during digital PCR analysis, each reaction well will contain an average of one template molecule (can be either the N or M allele). The digital PCR analysis will be carried out using a pair of primers, plus two TaqMan probes, one specific to the N allele, while the other one specific to the M allele. The number of wells which are positive only for M and the number of wells which are positive only for N will be counted. The ratio of these wells will be used to determine if there is evidence of allelic imbalance. Statistical evidence of allelic imbalance can be sought by methods well-known to those of skill in the art, such as using SPRT. In one variant of this analysis, it is also possible to count the number of wells which are positive either for M only or for M and N; as well as to count the number of wells which are positive either for N only or for M and N; and to derive a ratio of these counts. Once again, statistical evidence of allelic imbalance can be sought by methods well-known to those of skill in the art, such as using SPRT.

[0258] The dosage determination of fetal gene mutation, called digital relative mutant dosage (RMD), was validated using female/male (XX/XY) DNA mixtures. Blood cell DNA from a male and a female was each mixed with male DNA, producing samples with XX or XY genotypes in a background of XY at fractional concentrations of 25% and 50%, respectively, as shown in FIG. 24A.

[0259] In addition, blood cell samples were also obtained from 12 male and 12 female subjects. The female blood cell DNA (genotype XX) was each mixed with a 3-fold excess of male blood cell DNA (genotype XY), thus producing 12 DNA mixtures with 25% of DNA with XX genotype in a background of 75% DNA with XY genotype, with results shown in FIG. 24B.

[0260] An aim of the SPRT analysis was to determine the minor genotype present in the background DNA. In DNA mixtures with 25% of XX DNA in a background of 75% XY DNA, the minor allele would be the Y derived from the 75% of DNA. Since 25% of the DNA in the sample was of XX genotype, if there were a total of 200 molecules of DNA in the sample, then 150 molecules would have originated from the XY individual. Hence, the number of Y alleles would be expected to be 75. The number of X alleles contributed by the male proportion of DNA (genotype XY) is also 75. The number of X alleles contributed by the female (genotype XX) is 50 (2 times 25). Therefore, X to Y ratio is 125/75 = (1 +25%) / (1 – 25%) = 5/3.

[0261] For the second part of this study, blood cell samples were obtained from male and female subjects carrying HbE (G → A) and CD41/42 (CTTT/-) mutations on the *beta-globin* gene, i.e., the *hemoglobin, beta* (*HBB*) gene. To mimic maternal plasma samples obtained from heterozygous mothers (MN, where M= mutant and N= wildtype) bearing male fetuses with all possible genotypes (MM, MN or NN), blood cell DNA from males who were either homozygous for the wildtype alleles (NN) or heterozygous (MN) for one of the two mutations was each mixed with a blood cell DNA sample collected from females heterozygous for the same mutation (MN). DNA mixtures at various fractional male/mutant DNA concentrations were thus produced. Blood cell DNA sample from a female homozygous for the CD41/42 deletion (MM) was also used for preparing the DNA mixtures. To ensure an accurate male proportion for the SPRT classification, the fractional male DNA concentration of each DNA mixture was determined using the *ZFY/X* assay.

[0262] The digital *ZFY/X* assay was used for validating the SPRT as well as determining the fractional male DNA concentration in the DNA mixtures. The dosage of *Zinc Finger* protein sequences on chromosome X (*ZFX*) and Y (*ZFY*) was determined by digital PCR analysis. An 87-bp amplicon of the *ZFX* and *ZFY* loci was first co-amplified by the forward primer 5'-CAAGTGCTGGACTCAGATGTAACTG-3' and the reverse primer 5'-TGAAGTAATGTCAGAAGCTAAAACATCA-3'. Two chromosome-specific TaqMan probes were designed to distinguish between the chromosome X and Y paralogs, and their sequences were 5'-(VIC)TCTTTAGCACATTGCA(MGBNFQ)-3' and 5'-(FAM)TCTTTACCACACTGCAC(MGBNFQ)-3', respectively.

[0263] The mutant dosage in the DNA mixtures was determined by digital PCR analysis of the normal allele relative to the mutant allele. For the HbE mutation, an 87-bp amplicon of

the normal and mutant alleles was first co-amplified by the forward primer 5'-GGGCAAGGTGAACGTGGAT -3' and the reverse primer 5'-CTATTGGTCTCCTTAAACCTGTCTTGTAA -3'. Two allele-specific TaqMan probes were designed to distinguish between the normal (G) and mutant (A) alleles, and their sequences were 5'-(VIC) TTGGTGGTGAGGCC (MGBNFQ)-3' and 5'-(FAM) TTGGTGGTAAGGCC (MGBNFQ)-3', respectively. Results for the HbE mutation are shown in FIG. 25.

[0264] For the CD41/42 deletion mutation, an 87- and 83-bp amplicon of the normal and mutant alleles was first co-amplified by the forward primer 5'-TTTTCCCACCCTTAGGCTGC -3' and the reverse primer 5'-ACAGCATCAGGAGTGGACAGATC -3', respectively. Two allele-specific TaqMan probes were designed to distinguish between the normal (without deletion) and mutant (with deletion) alleles, and their sequences were 5'-(VIC) CAGAGGTTCTTTGAGTCCT (MGBNFQ)-3' and 5'-(FAM) AGAGGTTGAGTCCTT (MGBNFQ)-3', respectively. Results for the HbE mutation are shown in FIGS. 26A and 26B.

[0265] These experiments were carried out on the BioMark™ System (Fluidigm) using the 12.765 Digital Arrays (Fluidigm). The reaction for one panel was set up using 2X TaqMan Universal PCR Master Mix (Applied Biosystems) in a reaction volume of 10 μL. For the CD41/42 and *ZFY/X* assays, each reaction contained 1X TaqMan Universal PCR Master Mix, 900 nM of each primer, 125 nM of each probe and 3.5 μL of DNA mixture at 1 ng/μL. For the HbE assay, 250 nM and 125 nM of probes targeting the normal (G) and mutant (A) alleles were added, respectively. The sample/ assay mixture was loaded into the Digital Array by the NanoFlex™ IFC controller (Fluidigm). The reaction was carried out on the BioMark™ System for signal detection. The reaction was initiated at 50°C for 2 min, followed by 95°C for 10 min and 50 cycles of 95°C for 15s and 57°C (for *ZFY/X* and CD41/42) or 56°C (for HbE) for 1 min. At least one reaction panel was used for each case, and data were aggregated from extra panels for samples which remained unclassified until a decision could be made.

[0266] It will also be obvious to those of skill in the art that the digital PCR can be performed using methods well-known to those of skill in the art, e.g. microfluidics chips, nanoliter PCR microplate systems, emulsion PCR, polony PCR, rolling-circle amplification, primer extension and mass spectrometry.

IX. EXAMPLE WITH CANCER

[0267] In one embodiment, the present invention may be performed to classify a sample as having allelic ratio skewing or not, as may occur in a cancerous tumor. In one aspect, for each case, the number of wells with positive signal for the A allele only, the G allele only, and both alleles were determined by digital PCR. The reference allele was defined as the allele with the smaller number of positive wells. (In the unlikely scenario that both alleles have the same number of positive wells, then either can be used as the reference allele.) The inferred average concentration of the reference allele per well (m_r) was calculated using the total number of wells negative for the reference allele, irrespective whether the other allele was positive, according to the Poisson probability density function. We use a hypothetical example to illustrate the calculation.

[0268] In a 96-well reaction, 20 wells are positive for the A allele, 24 wells are positive for the G allele, and 28 wells are positive for both alleles. The A allele would be regarded as the reference allele because less wells are positive for this allele. The number of wells negative for the reference allele would be 96 - 20 - 28 = 48. Therefore, m_r can be calculated using the Poisson distribution and would be -1n(48/96) = 0.693.

[0269] In the context of LOH detection, the null hypothesis refers to a sample that is assumed to lack allelic ratio skewing caused by the presence of a deletion of one allele. Under this assumption, the expected ratio of the number of positive wells for the two alleles would be 1:1 and, thus, the expected proportion of informative wells (wells positive for only one allele) containing the potentially overrepresented allele would be 0.5.

[0270] In the context of LOH detection, the alternative hypothesis refers to a sample that is assumed to have allelic ratio skewing caused by the presence of a deletion of one allele in 50% of the cells of the sample. As the allelic ratio between the overrepresented allele and the reference allele is 2:1, the average concentration of the overrepresented allele per well would be doubled that of the reference allele. However, the number of wells positive for the overrepresented allele would not be simply two times that for the reference allele but would follow the Poisson distribution.

[0271] An informative well is defined as a well positive for either the A or the G allele but not for both alleles. The calculation of the expected proportion of the number of wells containing the overrepresented alleles for samples with allelic ratio skewing is the same as is shown in Table 600. In the above example, if LOH is present in 50% of tumor cells, the

average concentration of the G allele per well would be 2 times 0.693 = 1.386. If LOH is present in more than 50% of the tumor cells, then the average concentration of the G allele per well would be according to the formula: 1/[1-(proportion with LOH)] × m_r.

[0272] The expected proportion of wells positive for the G allele would be $1 - e^{-1..386} = 0.75$ (i.e., 75% or 72 wells). Assuming that the positivity of a well for the A or G allele is independent, 0.5 x 0.75 = 0.375 of the wells would be positive for both the A and G alleles. Hence, 0.5 – 0.375 = 0.125 of the wells would be positive for the A allele only and 0.75 – 0.375 = 0.375 of the wells would be positive for the G allele only. Therefore, the proportion of informative wells would be 0.125 + 0.375 = 0.5. The expected proportion of informative wells carrying the G allele would be 0.375/0.5 = 0.75. This expected value for P_r can then be used for the construction of appropriate SPRT curves for determining whether allelic ratio skewing (i.e. LOH in this context) is present in the sample.

[0273] The actual proportion of informative wells carrying the non-reference allele experimentallly determined by the digital PCR analysis (P_r) was then used to determine whether the null or alternative hypothesis would be accepted, or whether further analysis with more wells would be necessary. The decision boundaries for P_r to accept the null or alternative hypothesis was calculated based on a threshold likelihood ratio of 8 as this value had been shown to provide satisfactory performance to discriminate samples with and without allelic imbalance in the context of cancer detection (Zhou, W, et al. (2001) *Nat Biotechnol* 19, 78-81; Zhou et al 2002, supra). In the above example, the number of informative wells would be 20 +24 = 44 and the experimentally obtained P_r would be 24/44 = 0.5455. The decision boundaries would be ≤ 0.5879 to accept the null hypothesis and ≥ 0.6739 to accept the alternative hypothesis. Therefore, the sample in this example would be classified as NOT having allelic ratio skewing.

[0274] In conclusion, we outlined an approach to detect sequence imbalance in a sample. In one embodiment, this invention can be used for the noninvasive detection of fetal chromosomal aneuploidy, such as trisomy 21 by analysis of fetal nucleic acids in maternal plasma. This approach can also be applied to other biological materials containing fetal nucleic acids, including amniotic fluid, chorionic villus samples, maternal urine, endocervical samples, maternal saliva, etc. First, we demonstrated the use of this invention for determining allelic imbalance of a SNP on *PLAC4* mRNA, a placenta-expressed transcript on chromosome 21, in maternal plasma of women bearing trisomy 21 fetuses. Second, we

demonstrated that our invention can be used as a non-polymorphism based method, through relative chromosome dosage (RCD) analysis, for the noninvasive prenatal detection of trisomy 21. Such a digital RCD-based approach involves the direct assessment of whether the total copy number of chromosome 21 in a sample containing fetal DNA is overrepresented with respect to a reference chromosome. Even without elaborate instrumentation, digital RCD allows the detection of trisomy 21 in samples containing 25% fetal DNA. We applied the sequential probability ratio test (SPRT) to interpret the digital PCR data. Computer simulation analyses confirmed the high accuracy of the disease classification algorithm.

[0275] We further outlined that the approach can be applied to the determination of other forms of nucleic acid sequence imbalances other than chromosomal aneuploidy, such as that for the detection of fetal mutation or polymorphism detection in maternal plasma and regional gains and losses in the genomes of malignant cells through the analysis of tumor-derived nucleic acids in plasma.

[0276] Any of the software components or functions described in this application, may be implemented as software code to be executed by a processor using any suitable computer language such as, for example, Java, C++ or Perl using, for example, conventional or object-oriented techniques. The software code may be stored as a series of instructions, or commands on a computer readable medium for storage and/or transmission, suitable media include random access memory (RAM), a read only memory (ROM), a magnetic medium such as a hard-drive or a floppy disk, or an optical medium such as a compact disk (CD) or DVD (digital versatile disk), flash memory, and the like. The computer readable medium may be any combination of such storage or transmission devices.

[0277] Such programs may also be encoded and transmitted using carrier signals adapted for transmission via wired, optical, and/or wireless networks conforming to a variety of protocols, including the Internet. As such, a computer readable medium according to an embodiment of the present invention may be created using a data signal encoded with such programs. Computer readable media encoded with the program code may be packaged with a compatible device or provided separately from other devices (e.g., via Internet download). Any such computer readable medium may reside on or within a single computer program product (e.g. a hard drive or an entire computer system), and may be present on or within different computer program products within a system or network. A computer system may

include a monitor, printer, or other suitable display for providing any of the results mentioned herein to a user.

[0278] An example of a computer system is shown in FIG. 27. The subsystems shown in FIG. 27 are interconnected via a system bus 2775. Additional subsystems such as a printer 2774, keyboard 2778, fixed disk 2779, monitor 2776, which is coupled to display adapter 2782, and others are shown. Peripherals and input/output (I/O) devices, which couple to I/O controller 2771, can be connected to the computer system by any number of means known in the art, such as serial port 2777. For example, serial port 2777 or external interface 2781 can be used to connect the computer apparatus to a wide area network such as the Internet, a mouse input device, or a scanner. The interconnection via system bus allows the central processor 2773 to communicate with each subsystem and to control the execution of instructions from system memory 2772 or the fixed disk 2779, as well as the exchange of information between subsystems. The system memory 2772 and/or the fixed disk 2779 may embody a computer readable medium.

[0279] The above description of exemplary embodiments of the invention has been presented for the purposes of illustration and description. It is not intended to be exhaustive or to limit the invention to the precise form described, and many modifications and variations are possible in light of the teaching above. The embodiments were chosen and described in order to best explain the principles of the invention and its practical applications to thereby enable others skilled in the art to best utilize the invention in various embodiments and with various modifications as are suited to the particular use contemplated.

[0280] All publications, patents, and patent applications cited herein are hereby incorporated by reference in their entirety for all purposes.

WHAT IS CLAIMED IS:

1. A method for determining whether a nucleic acid sequence imbalance exists within a biological sample, the method comprising:

receiving data from a plurality of reactions, wherein the data includes:

(1) a first set of quantitative data indicating a first amount of a clinically relevant nucleic acid sequence; and

(2) a second set of quantitative data indicating a second amount of a background nucleic acid sequence different from the clinically relevant nucleic acid sequence;

determining a parameter from the two data sets;

deriving a first cutoff value from an average concentration of a reference nucleic acid sequence in each of the plurality of reactions, wherein the reference nucleic acid sequence is either the clinically relevant nucleic acid sequence or the background nucleic acid sequence;

comparing the parameter to the first cutoff value; and

based on the comparison, determining a classification of whether a nucleic acid sequence imbalance exists.

2. The method of claim 1 wherein the first set of data are obtained from one or more first markers that each detect a presence of a part of the clinically relevant nucleic acid sequence in a reaction, and wherein the second set of data are obtained from one or more second markers that each detect a presence of a part of the background nucleic acid sequence in a reaction.

3. The method of claim 1, further comprising:

determining the average concentration of the reference nucleic acid sequence in each of the plurality of reaction using an inverse of a probability distribution having an input of a value derived from the data for the reference nucleic acid sequence.

4. The method of claim 1 wherein the clinically relevant nucleic acid sequence is from chromosome 21 and the background nucleic acid sequence is from a chromosome other than chromosome 21.

5. The method of claim 1 wherein the clinically relevant nucleic acid sequence is from chromosome 18 or 13 and the background nucleic acid sequence is from a chromosome other than chromosome 18 or 13 respectively.

6. The method of claim 1 wherein the clinically relevant nucleic acid sequence is an allele of a genetic polymorphism, and the background nucleic acid sequence is another allele of the genetic polymorphism.

7. The method of claim 1 wherein the clinically relevant nucleic acid sequence is a mutated copy of the cystic fibrosis transmembrane conductance regulator (CFTR) gene, the beta-globin gene or the alpha-globin gene, and the background nucleic acid sequence is from the wildtype copy of the corresponding gene.

8. The method of claim 1 wherein the biological sample is plasma or serum from a pregnant woman.

9. The method of claim 1 wherein a reaction is an amplification reaction.

10. The method of claim 9 wherein a reaction is a part of a digital PCR process.

11. The method of claim 1 wherein a reaction is a sequencing reaction.

12. The method of claim 1 wherein first portions of the clinically relevant nucleic acid sequence and the background nucleic acid sequence are from a first individual and second portions of the clinically relevant nucleic acid sequence and the background nucleic acid sequence are from a second individual.

13. The method of claim 12 wherein the cutoff value is based on a measurement of one of the first portions or a measurement of one of the second portions.

14. The method of claim 1, further comprising comparing the parameter to a second cutoff value.

15. The method of claim 14 wherein the classifications include disease state, non-disease state, and non-classifiable.

16. The method of claim 14 wherein the classifications include homozygous, heterozygous, and non-classifiable.

17. The method of claim 14 wherein the second cutoff value is based on a ratio of the first amount of the clinically relevant nucleic acid sequence relative to the second amount of the background nucleic acid sequence in a non-diseased state.

18. The method of claim 1 wherein the parameter is a ratio of the first amount of the clinically relevant nucleic acid sequence relative to the second amount of the background nucleic acid sequence.

19. The method of claim 1 wherein calculating the first cutoff value includes using at least one of: sequential probability ratio testing, false discovery rates, confidence intervals, and receiver operating characteristic curves.

20. The method of claim 1 wherein deriving the first cutoff value includes:

determining a proportion P1 of informative reactions containing an overrepresented nucleic acid sequence, which is either the reference or non-reference nucleic acid sequence; and

calculating the first cutoff value from the first proportion P1.

21. The method of claim 20 wherein determining the proportion P1 comprises:

determining a first probability of a reaction containing at least one of the overrepresented nucleic acid sequence;

calculating a second probability that a reaction is informative; and

calculating the proportion P1 using the first probability and the second probability.

22. The method of claim 21 wherein the first probability is determined by multiplying the average concentration of the reference nucleic acid sequence by an expected ratio relative to the non-reference nucleic acid sequence.

23. The method of claim 21 wherein the first probability is determined using the Poisson distribution having the average concentration of the overrepresented nucleic acid sequence in each of the plurality of reactions as an input.

24. The method of claim 21, further comprising:

determining a third probability of a reaction containing at least one of the underrepresented nucleic acid sequence, wherein calculating the second probability that the reaction is informative includes assuming that first probability and the second probability are independent.

25. A method for determining whether a nucleic acid sequence imbalance exists within a biological sample, the method comprising:

receiving data from a plurality of reactions, wherein the data includes:

(1) a first set of quantitative data indicating a first amount of a clinically relevant nucleic acid sequence; and

(2) a second set of quantitative data indicating a second amount of a background nucleic acid sequence different from the clinically relevant nucleic acid sequence, wherein the clinically relevant nucleic acid sequence and the background nucleic acid sequence come from a first type of cells and from one or more second types of cells;

determining a parameter from the two data sets;

deriving a first cutoff value from a first percentage resulting from a measurement of an amount of a nucleic acid sequence from the first type of cells in the biological sample;

comparing the parameter to the first cutoff value; and

based on the comparison, determining a classification of whether a nucleic acid sequence imbalance exists.

26. The method of claim 25 wherein the first type of cells is from a first organism and the second types of cells are from a second organism.

27. The method of claim 25 wherein deriving the first cutoff value includes:

determining a first average concentration of a reference nucleic acid sequence per reaction, wherein the reference nucleic acid sequence is the clinically relevant nucleic acid sequence or the background nucleic acid sequence that is underrepresented; and

multiplying the first average concentration by a factor derived from the first percentage to obtain a second average concentration of the nucleic acid sequence that is not the reference nucleic acid sequence.

28. The method of claim 27, further comprising:

determining the average concentration of the reference nucleic acid sequence in each of the plurality of reactions using an inverse of a probability distribution having an input of a value derived from the data for the reference nucleic acid sequence.

29. The method of claim 28 wherein the probability distribution is the Poisson distribution.

30. The method of claim 25 wherein the percentage is measured by determining an amount of a fetal-specific marker using quantitative real-time PCR, digital PCR, semiquantitative competitive PCR, real-competitive PCR, or mass spectrometry.

31. A computer program product comprising a computer readable medium encoded with a plurality of instructions for controlling a computing system to perform an operation for determining whether a nucleic acid sequence imbalance exists within a biological sample, the operation comprising the steps of:

receiving data from a plurality of reactions, wherein the data includes:

(1) a first set of quantitative data indicating a first amount of a clinically relevant nucleic acid sequence; and

(2) a second set of quantitative data indicating a second amount of a background nucleic acid sequence different from the clinically relevant nucleic acid sequence;

determining a parameter from the two data sets;

deriving a first cutoff value from an average concentration of a reference nucleic acid sequence in each of the plurality of reactions, wherein the reference nucleic acid sequence is either the clinically relevant nucleic acid sequence or the background nucleic acid sequence;

comparing the parameter to the first cutoff value; and

based on the comparison, determining a classification of whether a nucleic acid sequence imbalance exists.

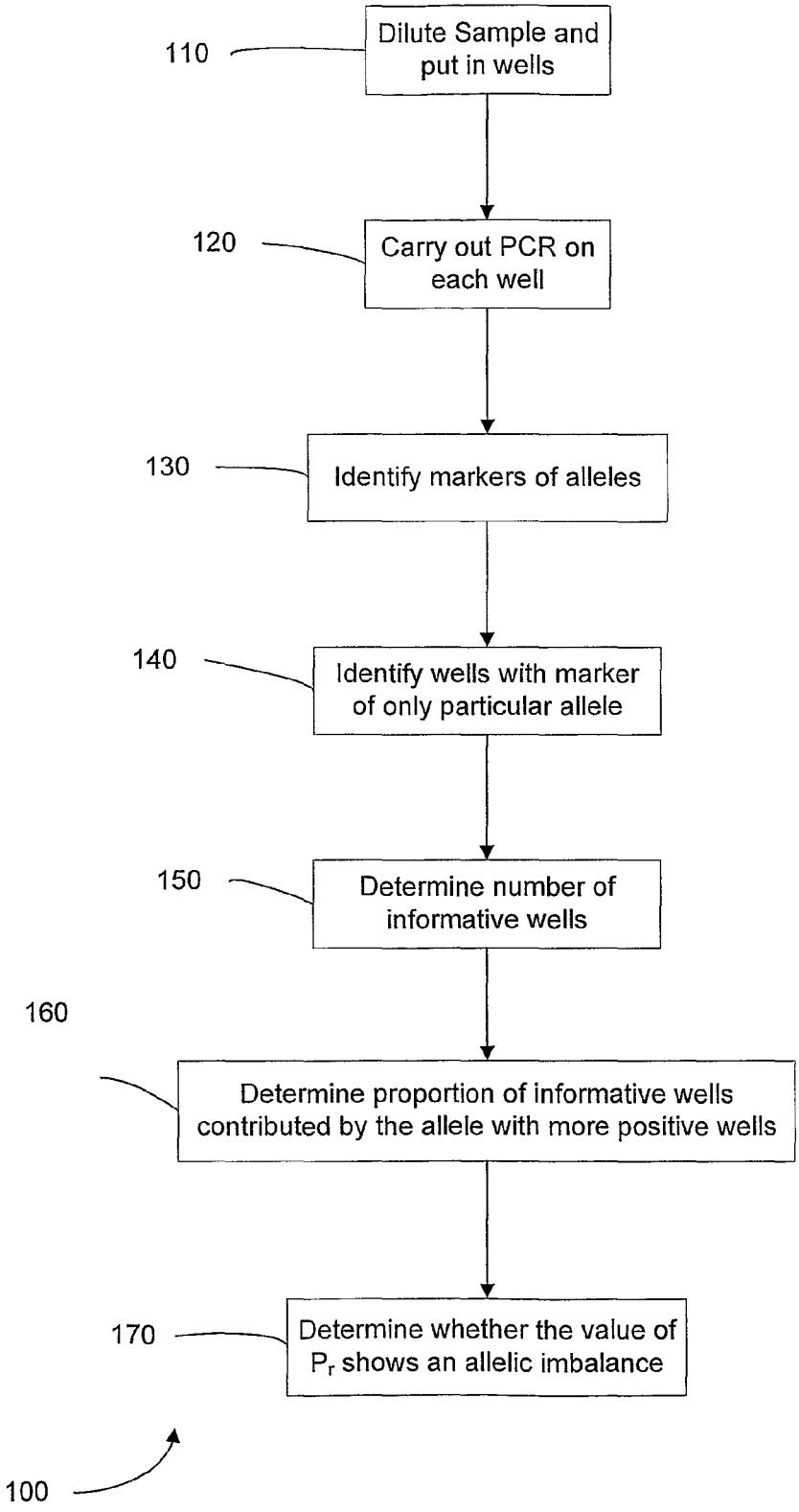

110
Dilute Sample and put in wells
120
Carry out PCR on each well
130
Identify markers of alleles
140
Identify wells with marker of only particular allele
150
Determine number of informative wells
160
Determine proportion of informative wells contributed by the allele with more positive wells
170
Determine whether the value of P_r shows an allelic imbalance
100

FIG. 1

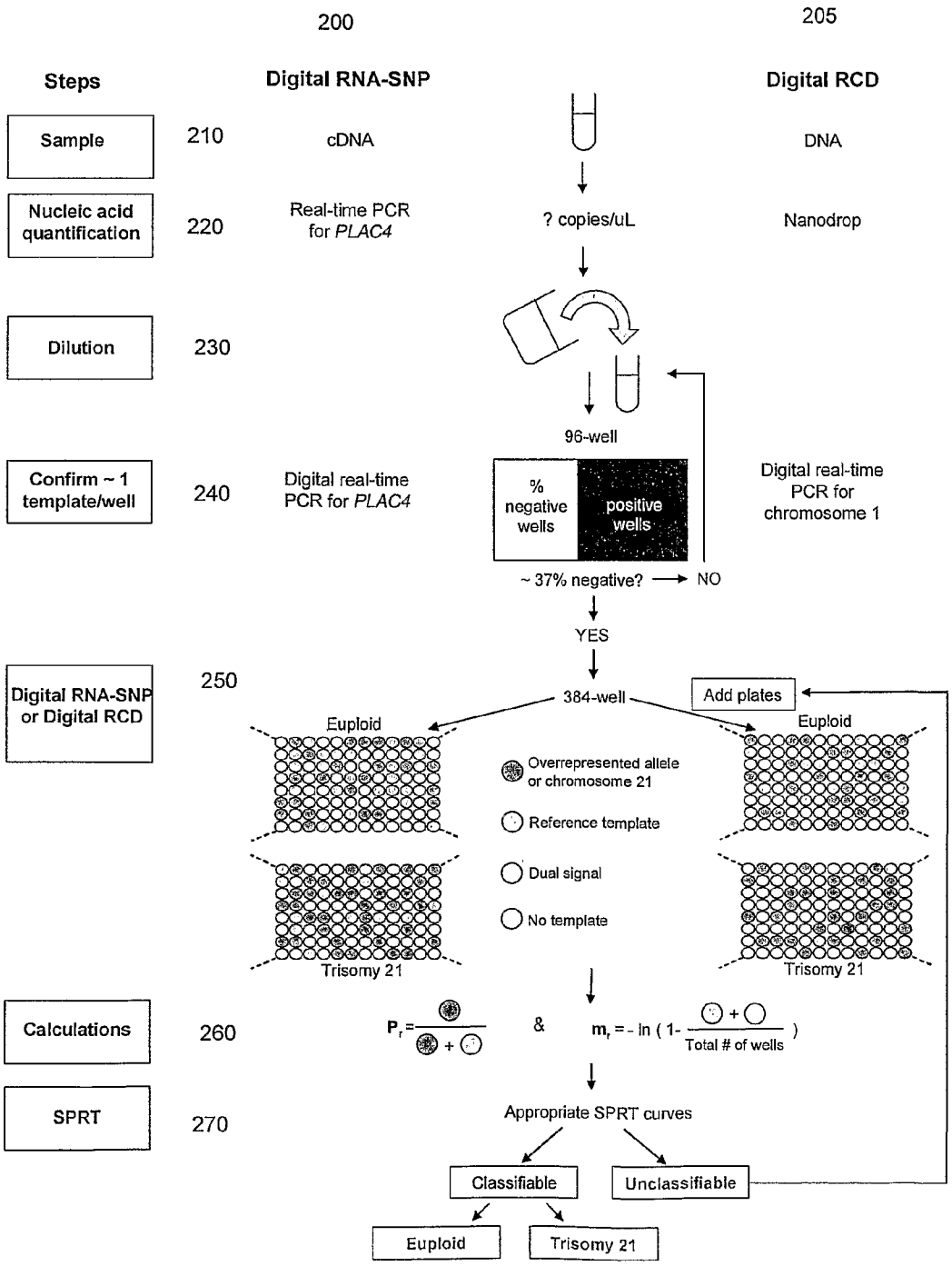

200
205
Steps
Digital RNA-SNP
Digital RCD
Sample
210
cDNA
DNA
Nucleic acid quantification
220
Real-time PCR for PLAC4
? copies/uL
Nanodrop
Dilution
230
96-well
Confirm ~ 1 template/well
240
Digital real-time PCR for PLAC4
% negative wells
positive wells
Digital real-time PCR for chromosome 1
~ 37% negative?
NO
YES
Digital RNA-SNP or Digital RCD
250
384-well
Add plates
Euploid
Euploid
Overrepresented allele or chromosome 21
Reference template
Dual signal
No template
Trisomy 21
Trisomy 21
Calculations
260
$P_r = \frac{\text{Overrepresented}}{\text{Overrepresented} + \text{Reference}}$
&
$m_r = -\ln\left(1 - \frac{\text{Reference} + \text{Dual}}{\text{Total \# of wells}}\right)$
SPRT
270
Appropriate SPRT curves
Classifiable
Unclassifiable
Euploid
Trisomy 21

FIG. 2A

Cancer	Common chromosomal aberrations	
	Gain	Loss
Bladder cancer	7	9p; 11q; 17p
Breast cancer	1q; 8q; 17q; 20q	8p; 13q; 16q; 18q
Colorectal cancer	7; 8q; 13q; 20q	1p; 5q; 8p; 15q; 17p; 18q
Hepatocellular carcinoma	1q; 8q; 20q	4q; 8p; 13q; 16q; 17p
Non - small cell lung carcinoma	1q; 3q; 5p; 8q	3p; 8p; 9p; 13q; 17p
Small cell lung cancinoma	3q; 5p; 8q; 19q	3p; 4q; 5q; 10q; 13q; 16q
Nasopharyngeal carcinoma	1q; 3q; 8; 12; 19	1p; 3p; 9; 11q; 13q; 14q; 16q
Prostate cancer	7; 8q; Xq	8p; 13q; 16q

FIG. 2B

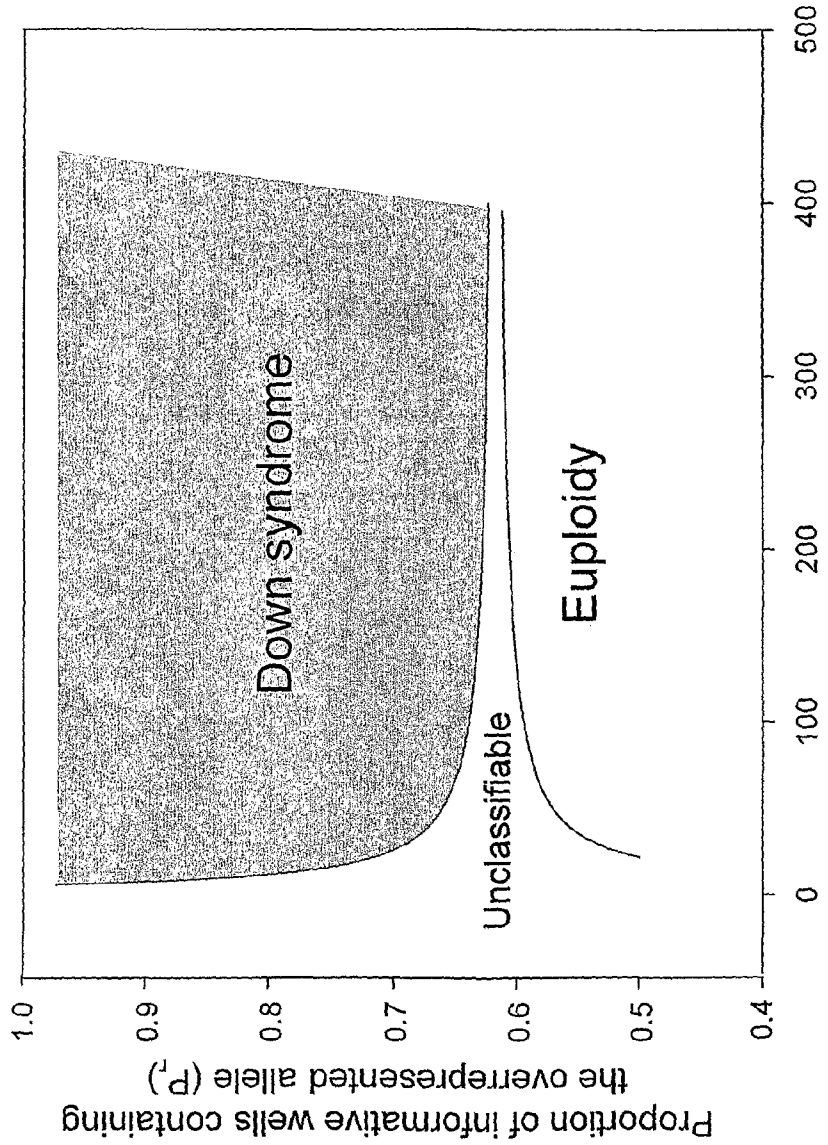

Down syndrome
Unclassifiable
Euploidy
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0
100
200
300
400
500
No. of informative wells
Proportion of informative wells containing the overrepresented allele (P_r)

FIG. 3

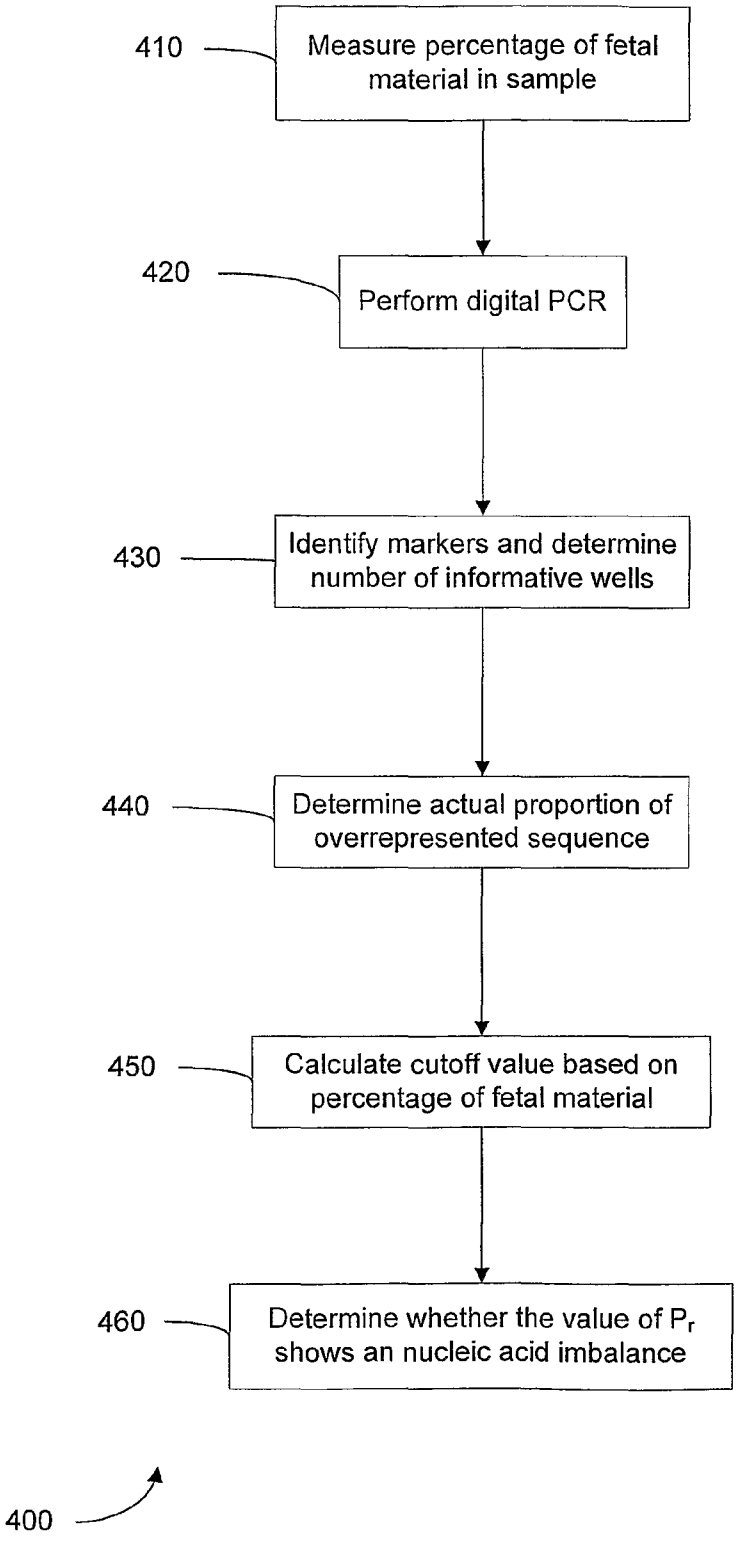

410
Measure percentage of fetal material in sample
420
Perform digital PCR
430
Identify markers and determine number of informative wells
440
Determine actual proportion of overrepresented sequence
450
Calculate cutoff value based on percentage of fetal material
460
Determine whether the value of P_r shows an nucleic acid imbalance
400

FIG. 4

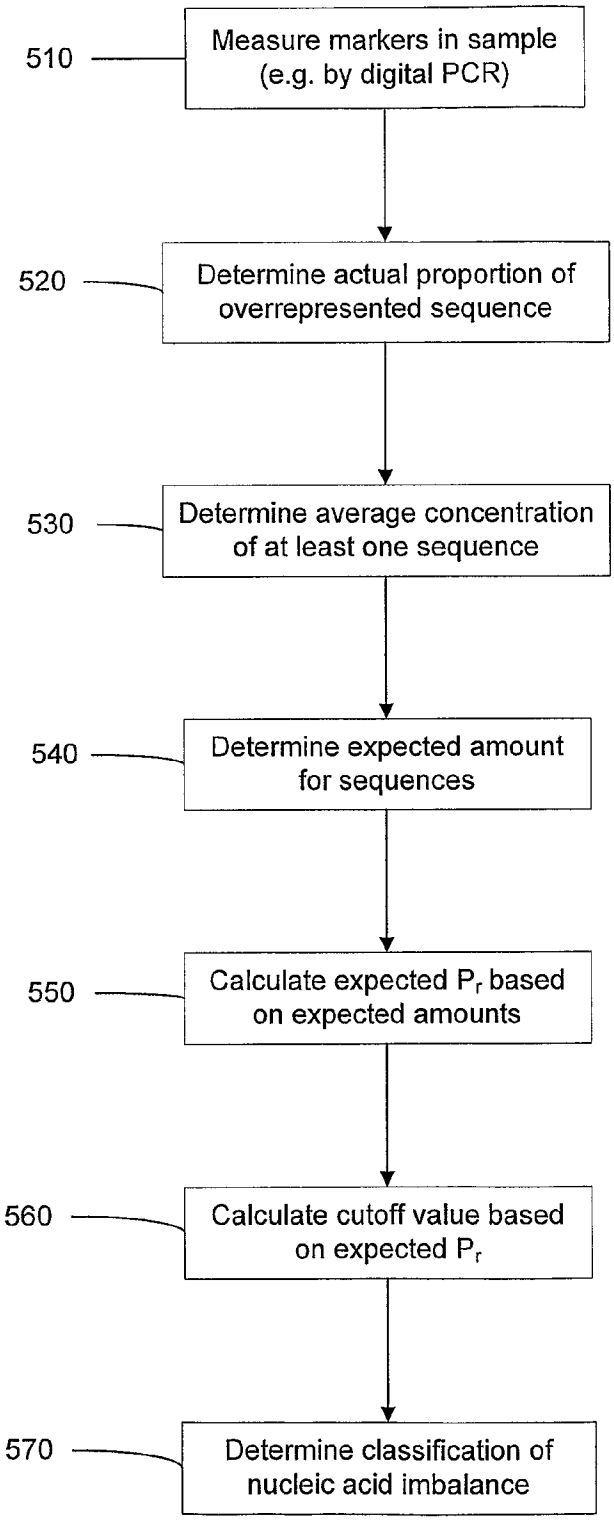

510
Measure markers in sample (e.g. by digital PCR)
520
Determine actual proportion of overrepresented sequence
530
Determine average concentration of at least one sequence
540
Determine expected amount for sequences
550
Calculate expected P_r based on expected amounts
560
Calculate cutoff value based on expected P_r
570
Determine classification of nucleic acid imbalance
500

FIG. 5

m_r	Proportion of wells positive for the reference allele	Proportion of wells positive for the overrepresented allele	Proportion of wells positive for both alleles	Proportion of wells positive for the reference allele ONLY	Proportion of wells positive for the overrepresented allele ONLY	Digital RNA-SNP allelic ratio	Proportion of informative wells	P_r
0.1	0.0952	0.1813	0.0173	0.0779	0.164	2.11	0.2419	0.68
0.2	0.1813	0.3297	0.0598	0.1215	0.2699	2.22	0.3914	0.69
0.3	0.2592	0.4512	0.1169	0.1422	0.3342	2.35	0.4765	0.7
0.4	0.3297	0.5507	0.1815	0.1481	0.3691	2.49	0.5173	0.71
0.5	0.3935	0.6321	0.2487	0.1448	0.3834	2.65	0.5282	0.73
0.6	0.4512	0.6988	0.3153	0.1359	0.3835	2.82	0.5194	0.74
0.7	0.5034	0.7534	0.3793	0.1241	0.3741	3.01	0.4983	0.75
0.8	0.5507	0.7981	0.4395	0.1112	0.3586	3.23	0.4698	0.76
0.9	0.5934	0.8347	0.4953	0.0981	0.3394	3.46	0.4375	0.78
1	0.6321	0.8647	0.5466	0.0855	0.3181	3.72	0.4036	0.79
1.1	0.6671	0.8892	0.5932	0.0739	0.296	4	0.3699	0.8
1.2	0.6988	0.9093	0.6354	0.0634	0.2739	4.32	0.3373	0.81
1.3	0.7275	0.9257	0.6734	0.054	0.2523	4.67	0.3063	0.82
1.4	0.7534	0.9392	0.7076	0.0458	0.2316	5.06	0.2774	0.83
1.5	0.7769	0.9502	0.7382	0.0387	0.212	5.48	0.2507	0.85
1.6	0.7981	0.9592	0.7656	0.0325	0.1937	5.95	0.2262	0.86
1.7	0.8173	0.9666	0.79	0.0273	0.1766	6.47	0.2039	0.87
1.8	0.8347	0.9727	0.8119	0.0228	0.1608	7.05	0.1836	0.88
1.9	0.8504	0.9776	0.8314	0.019	0.1462	7.69	0.1652	0.88
2	0.8647	0.9817	0.8488	0.0158	0.1329	8.39	0.1487	0.89

600

FIG. 6

m_r	Percentage of fetal DNA in the sample									
	10%		25%		50%		100%			
	Percentage of informative wells	P_r	Percentage of informative wells	P_r	Percentage of informative wells	P_r	Percentage of informative wells	P_r		
0.1	18%	0.51	18%	0.53	19%	0.56	21%	0.61		
0.2	30%	0.51	31%	0.53	32%	0.56	35%	0.61		
0.3	39%	0.51	40%	0.53	41%	0.57	43%	0.62		
0.4	45%	0.51	45%	0.54	46%	0.57	48%	0.63		
0.5	48%	0.52	49%	0.54	49%	0.57	51%	0.63		
0.6	50%	0.52	50%	0.54	50%	0.58	51%	0.64		
0.7	50%	0.52	50%	0.54	50%	0.58	50%	0.65		
0.8	49%	0.52	49%	0.54	49%	0.58	48%	0.65		
0.9	48%	0.52	47%	0.55	47%	0.59	46%	0.66		
1	46%	0.52	45%	0.55	44%	0.59	43%	0.67		
1.1	44%	0.52	43%	0.55	42%	0.6	40%	0.68		
1.2	41%	0.52	40%	0.55	39%	0.6	37%	0.69		
1.3	39%	0.52	38%	0.55	36%	0.6	34%	0.69		
1.4	36%	0.52	35%	0.56	33%	0.61	31%	0.7		
1.5	34%	0.52	33%	0.56	31%	0.61	28%	0.71		
1.6	31%	0.52	30%	0.56	28%	0.62	26%	0.72		
1.7	29%	0.53	28%	0.56	26%	0.62	23%	0.73		
1.8	27%	0.53	25%	0.57	24%	0.63	21%	0.73		
1.9	24%	0.53	23%	0.57	21%	0.63	19%	0.74		
2	22%	0.53	21%	0.57	20%	0.64	17%	0.75		

700

FIG. 7

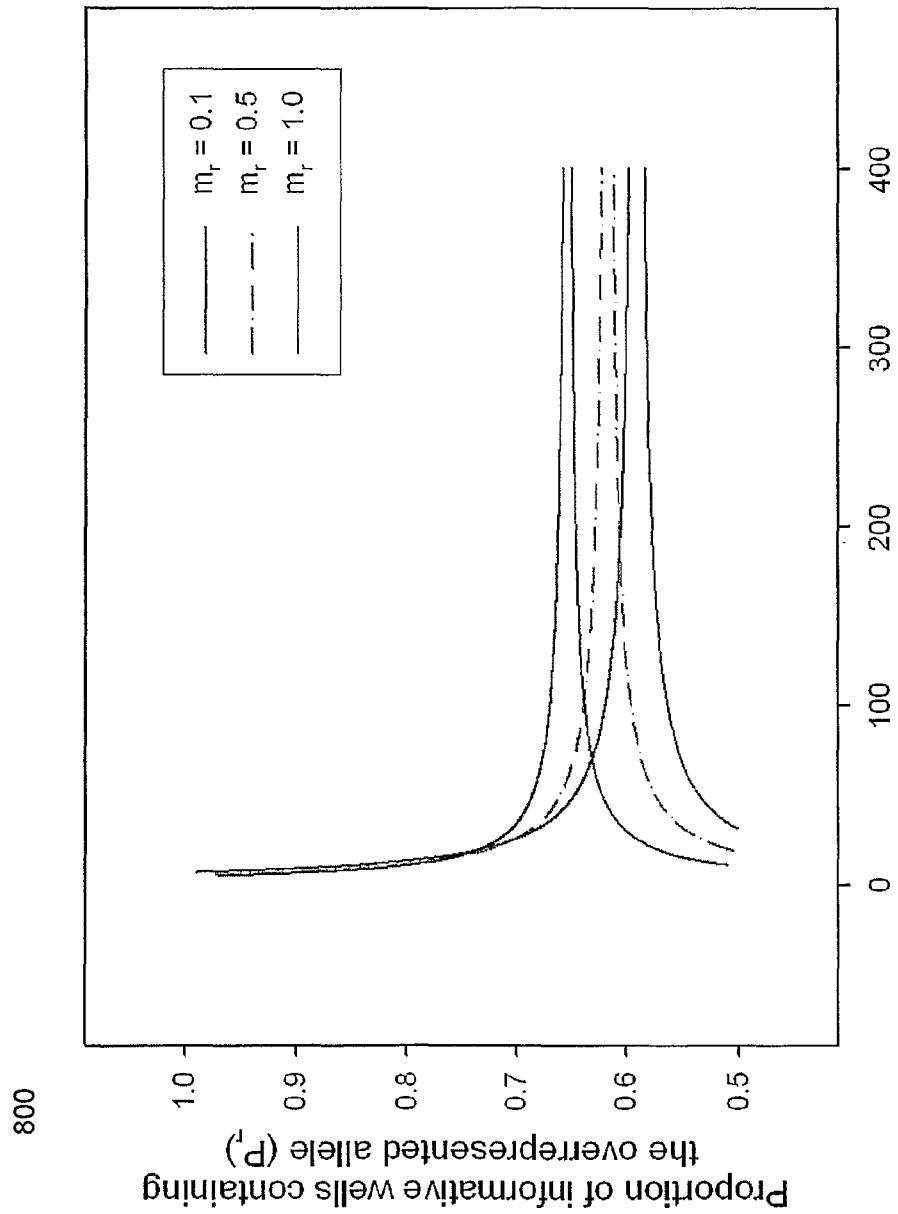

Proportion of informative wells containing the overrepresented allele (P_i)
1.0
0.9
0.8
0.7
0.6
0.5
0
100
200
300
400
No. of informative wells
m_r = 0.1
m_r = 0.5
m_r = 1.0
800

FIG. 8

m_F	Classification	New algorithm		Old algorithm		m_F	Classification	New algorithm		Old algorithm	
		Euploid	Trisomy 21	Euploid	Trisomy 21			Euploid	Trisomy 21	Euploid	Trisomy 21
0.1	Correct	0%	37%	0%	35%	1.1	correct	89%	92%	14%	35%
	Incorrect	3%	2%	3%	1%		Incorrect	1%	0%	4%	0%
	Unclassified	97%	61%	97%	63%		Unclassified	10%	7%	81%	65%
0.2	Correct	6%	64%	0%	40%	1.2	correct	90%	92%	11%	35%
	Incorrect	4%	2%	4%	0%		Incorrect	1%	0%	4%	0%
	Unclassified	90%	34%	96%	60%		Unclassified	9%	8%	84%	65%
0.3	Correct	34%	75%	6%	39%	1.3	correct	91%	92%	7%	34%
	Incorrect	4%	2%	4%	0%		Incorrect	1%	1%	4%	0%
	Unclassified	62%	24%	90%	61%		Unclassified	8%	7%	89%	66%
0.4	Correct	54%	81%	14%	39%	1.4	correct	92%	91%	5%	34%
	Incorrect	3%	1%	4%	0%		Incorrect	1%	1%	4%	0%
	Unclassified	43%	19%	82%	61%		Unclassified	8%	9%	91%	66%
0.5	Correct	64%	85%	20%	38%	1.5	correct	92%	90%	2%	32%
	Incorrect	3%	1%	4%	0%		Incorrect	1%	1%	4%	0%
	Unclassified	33%	14%	76%	62%		Unclassified	8%	9%	93%	68%
0.6	Correct	74%	88%	23%	38%	1.6	correct	92%	89%	1%	31%
	Incorrect	2%	1%	4%	0%		Incorrect	1%	1%	4%	0%
	Unclassified	24%	11%	72%	62%		Unclassified	8%	10%	94%	69%
0.7	Correct	80%	89%	25%	37%	1.7	correct	92%	89%	1%	30%
	Incorrect	2%	1%	4%	0%		Incorrect	1%	1%	4%	0%
	Unclassified	18%	10%	71%	63%		Unclassified	7%	10%	96%	70%
0.8	Correct	83%	91%	23%	37%	1.8	correct	92%	88%	0%	28%
	Incorrect	2%	1%	4%	0%		Incorrect	1%	1%	4%	0%
	Unclassified	15%	9%	73%	63%		Unclassified	7%	11%	96%	72%
0.9	Correct	86%	91%	21%	36%	1.9	correct	92%	87%	0%	26%
	Incorrect	1%	1%	4%	0%		Incorrect	1%	1%	3%	0%
	Unclassified	12%	8%	75%	64%		Unclassified	8%	12%	97%	74%
1	Correct	88%	92%	19%	36%	2	correct	91%	86%	0%	24%
	Incorrect	1%	0%	4%	0%		Incorrect	1%	1%	4%	0%
	Unclassified	11%	8%	77%	64%		Unclassified	9%	13%	96%	76%

900

FIG. 9A

m_r	Classification	New algorithm		Old algorithm		m_r	Classification	New algorithm		Old algorithm	
		Euploid	Trisomy 21	Euploid	Trisomy 21			Euploid	Trisomy 21	Euploid	Trisomy 21
0.1	correct	57%	88%	48%	89%	1.1	correct	100%	100%	92%	90%
	Incorrect	3%	1%	3%	1%		Incorrect	0%	0%	1%	0%
	Unclassified	40%	11%	49%	10%		Unclassified	0%	0%	7%	10%
0.2	correct	89%	98%	79%	90%	1.2	correct	100%	100%	90%	90%
	Incorrect	1%	0%	2%	0%		Incorrect	0%	0%	1%	0%
	Unclassified	10%	2%	19%	10%		Unclassified	0%	0%	9%	10%
0.3	correct	96%	99%	88%	90%	1.3	correct	100%	100%	89%	90%
	Incorrect	0%	0%	1%	0%		Incorrect	0%	0%	1%	0%
	Unclassified	4%	1%	11%	10%		Unclassified	0%	0%	10%	10%
0.4	correct	99%	100%	92%	90%	1.4	correct	100%	100%	87%	90%
	Incorrect	0%	0%	1%	0%		Incorrect	0%	0%	1%	0%
	Unclassified	1%	0%	7%	10%		Unclassified	0%	0%	12%	10%
0.5	correct	100%	100%	94%	90%	1.5	correct	100%	100%	85%	90%
	Incorrect	0%	0%	1%	0%		Incorrect	0%	0%	1%	0%
	Unclassified	0%	0%	6%	10%		Unclassified	0%	0%	13%	10%
0.6	correct	100%	100%	94%	90%	1.6	correct	100%	100%	82%	90%
	Incorrect	0%	0%	1%	0%		Incorrect	0%	0%	1%	0%
	Unclassified	0%	0%	5%	10%		Unclassified	0%	0%	17%	10%
0.7	correct	100%	100%	95%	90%	1.7	correct	100%	100%	79%	90%
	Incorrect	0%	0%	1%	0%		Incorrect	0%	0%	2%	0%
	Unclassified	0%	0%	5%	10%		Unclassified	0%	0%	19%	10%
0.8	correct	100%	100%	95%	90%	1.8	correct	100%	100%	74%	90%
	Incorrect	0%	0%	1%	0%		Incorrect	0%	0%	2%	0%
	Unclassified	0%	0%	5%	10%		Unclassified	0%	0%	24%	10%
0.9	correct	100%	100%	94%	90%	1.9	correct	100%	100%	70%	90%
	Incorrect	0%	0%	1%	0%		Incorrect	0%	0%	2%	0%
	Unclassified	0%	0%	6%	10%		Unclassified	0%	0%	28%	10%
1	correct	100%	100%	94%	90%	2	correct	100%	100%	67%	90%
	Incorrect	0%	0%	1%	0%		Incorrect	0%	0%	3%	0%
	Unclassified	0%	0%	6%	10%		Unclassified	0%	0%	30%	10%

950

FIG. 9B

m_T	Classification result	*Euploid*	*T21*	m_T	Classification result	*Euploid*	*T21*
0.1	Correct	57%	88%	0.9	Correct	100%	100%
	Incorrect	3%	1%		Incorrect	0%	0%
	Unclassified	40%	11%		Unclassified	0%	0%
0.2	Correct	89%	98%	1	Correct	100%	100%
	Incorrect	1%	0%		Incorrect	0%	0%
	Unclassified	10%	2%		Unclassified	0%	0%
0.3	Correct	96%	99%	1.1	Correct	100%	100%
	Incorrect	0%	0%		Incorrect	0%	0%
	Unclassified	4%	1%		Unclassified	0%	0%
0.4	Correct	99%	100%	1.2	Correct	100%	100%
	Incorrect	0%	0%		Incorrect	0%	0%
	Unclassified	1%	0%		Unclassified	0%	0%
0.5	Correct	100%	100%	1.3	Correct	100%	100%
	Incorrect	0%	0%		Incorrect	0%	0%
	Unclassified	0%	0%		Unclassified	0%	0%
0.6	Correct	100%	100%	1.4	Correct	100%	100%
	Incorrect	0%	0%		Incorrect	0%	0%
	Unclassified	0%	0%		Unclassified	0%	0%
0.7	Correct	100%	100%	1.5	Correct	100%	100%
	Incorrect	0%	0%		Incorrect	0%	0%
	Unclassified	0%	0%		Unclassified	0%	0%
0.8	Correct	100%	100%	2	Correct	100%	100%
	Incorrect	0%	0%		Incorrect	0%	0%
	Unclassified	0%	0%		Unclassified	0%	0%

1000

FIG. 10

m_r	Classification result	*Euploid*	*Trisomy 21*	m_r	Classification result	*Euploid*	*Trisomy 21*
0.1	Correct	38%	44%	0.7	Correct	95%	94%
	Incorrect	2%	2%		Incorrect	0%	1%
	Unclassified	60%	54%		Unclassified	5%	5%
0.2	Correct	64%	69%	0.8	Correct	95%	95%
	Incorrect	2%	2%		Incorrect	1%	0%
	Unclassified	34%	29%		Unclassified	4%	5%
0.3	Correct	78%	82%	0.9	Correct	96%	96%
	Incorrect	2%	1%		Incorrect	0%	0%
	Unclassified	20%	17%		Unclassified	4%	4%
0.4	Correct	85%	87%	1	Correct	97%	96%
	Incorrect	1%	1%		Incorrect	0%	0%
	Unclassified	14%	12%		Unclassified	3%	4%
0.5	Correct	89%	90%	1.5	Correct	97%	96%
	Incorrect	1%	1%		Incorrect	0%	0%
	Unclassified	10%	9%		Unclassified	3%	4%
0.6	Correct	92%	93%	2	Correct	96%	95%
	Incorrect	1%	1%		Incorrect	0%	0%
	Unclassified	7%	7%		Unclassified	4%	5%

1100

FIG. 11

Number of wells	Classification results	Fetal DNA percentage in the sample							
		10%		25%		50%		100%	
		Euploid	*Trisomy 21*	*Euploid*	*Trisomy 21*	*Euploid*	*Trisomy 21*	*Euploid*	*Trisomy 21*
384	Correct	0%	0%	6%	7%	47%	49%	89%	90%
	Incorrect	0%	0%	0%	1%	2%	2%	1%	1%
	Unclassified	100%	100%	93%	92%	51%	49%	10%	9%
768	Correct	0%	0%	23%	24%	73%	76%	98%	99%
	Incorrect	0%	0%	2%	2%	2%	1%	0%	0%
	Unclassified	100%	100%	75%	74%	25%	23%	2%	1%
1152	Correct	1%	1%	39%	40%	86%	87%	100%	100%
	Incorrect	0%	0%	2%	2%	1%	1%	0%	0%
	Unclassified	99%	99%	59%	58%	13%	12%	0%	0%
1536	Correct	2%	2%	50%	51%	93%	93%	100%	100%
	Incorrect	0%	0%	2%	2%	1%	1%	0%	0%
	Unclassified	98%	97%	48%	47%	7%	6%	0%	0%
1920	Correct	4%	4%	60%	60%	96%	96%	100%	100%
	Incorrect	0%	1%	2%	2%	0%	0%	0%	0%
	Unclassified	96%	95%	38%	37%	4%	3%	0%	0%
3840	Correct	18%	19%	84%	83%	100%	100%	100%	100%
	Incorrect	1%	1%	1%	1%	0%	0%	0%	0%
	Unclassified	81%	80%	15%	15%	18%	22%	0%	0%
7680	Correct	43%	45%	97%	97%	100%	100%	100%	100%
	Incorrect	2%	2%	0%	0%	0%	0%	0%	0%
	Unclassified	55%	53%	3%	3%	0%	0%	0%	0%

1200

FIG. 12

sample	geno-type[a]	no. of wells	no. of wells positive for individual alleles				m_r[b]	SPRT result		
			A only	G only	AG	all negative		P_r[c]	unclassifiable region[d]	classification[e]
Placental DNA										
N677	AG	384	85	83	126	90	0.79	0.51	0.63 - 0.65	euploid
N710	AG	384	102	83	73	126	0.52	0.55	0.61 - 0.63	euploid
N435	AGG	384	49	157	130	48	0.63	0.76	0.62 - 0.64	T21
N981	AAG	384	135	69	82	98	0.5	0.66	0.61 - 0.63	T21
Placental RNA										
V533	AG	384	103	93	71	117	0.56	0.53	0.61 - 0.63	euploid
V943	AG	384	89	100	74	121	0.55	0.53	0.61 - 0.63	euploid
N435	AGG	384	52	138	95	99	0.48	0.73	0.61 - 0.63	T21
T215	AAG	384	146	58	138	42	0.71	0.72	0.62 - 0.64	T21

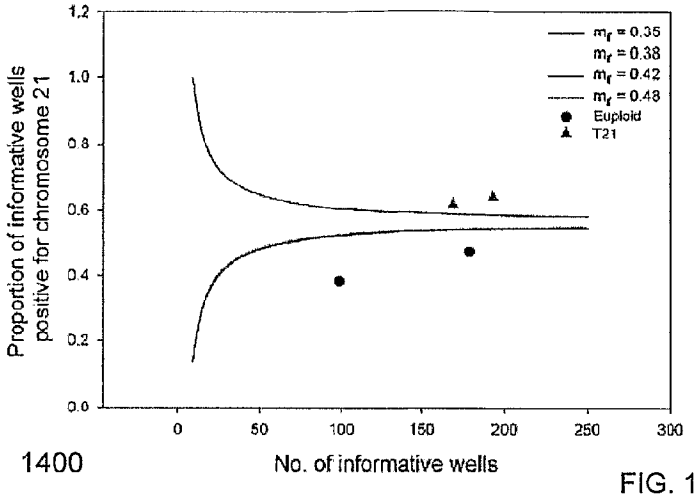
1300

FIG. 13A

sample	geno-type	no. of wells	no. of wells positive for individual alleles				m	SPRT result		
			A only	G only	AG	all negative		ratio	unclassifiable region	classification
M2390P	AG	384	90	100	97	97	0.67	0.526	0.62 - 0.64	euploid
M2391P	AG	384	97	105	65	117	0.55	0.52	0.61 - 0.63	euploid
M2473P	AG	384	66	92	34	192	0.3	0.582	0.59 - 0.62	euploid
M2524P	AG	384	29	28	3	324	0.08	0.509	0.54 - 0.64	euploid
M2528P	AG	384	112	85	44	143	0.41	0.569	0.60 - 0.62	euploid
M2601P	AG	384	90	101	72	121	0.55	0.529	0.61 - 0.63	euploid
M2607P	AG	384	73	91	57	163	0.41	0.555	0.60 - 0.63	euploid
M2638P	AG	384	66	90	52	176	0.37	0.577	0.59 - 0.62	euploid
M2639P	AG	384	71	56	17	240	0.21	0.559	0.58 - 0.62	euploid
M2525P	AAG	384	110	53	21	200	0.21	0.675	0.58 - 0.61	T21
M2272P	AAG	768	246	127	112	283	0.37	0.66	0.60 - 0.61	T21
M2718P	AGG	384	66	114	66	138	0.42	0.633	0.60 - 0.62	T21
M1519P	AGG	384	58	130	54	142	0.34	0.691	0.59 - 0.62	T21

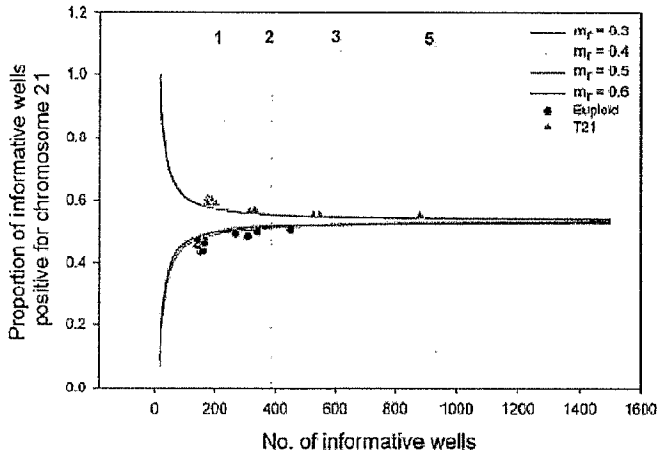
1350

FIG. 13B



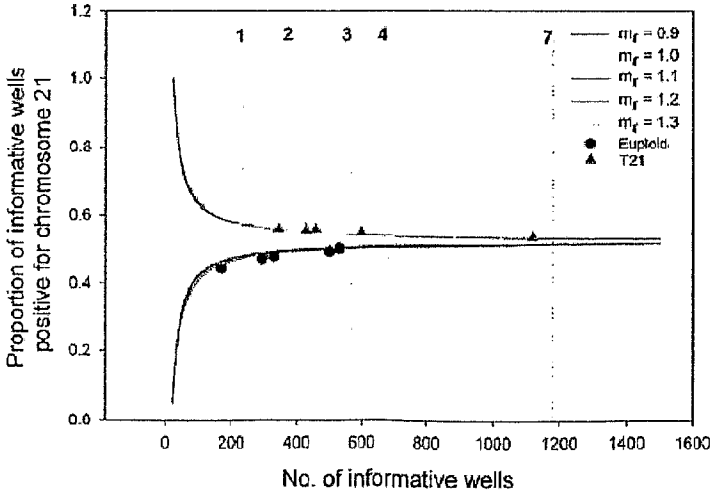
1400

FIG. 14A



1430

FIG. 14B



1460

FIG. 14C

sample	genotype	no. of wells	no. of wells positive for individual alleles				m_r[a]	SPRT result		
			A only	G only	GA	all negative		P_r[b]	unclassifiable region	classification[c]
V943	AG	765	168	209	117	271	0.47	0.554	0.609- 0.621	euploid
N435	AGG	765	92	281	105	287	0.30	0.753	0.597- 0.610	T21
T21-5	AAG	765	267	126	141	231	0.43	0.679	0.607- 0.618	T21

[a] no. of reference molecules per well

[b] P_r values were calculated by the equation:

$$\frac{\text{no. of wells positive for the overrepresented allele}}{\text{no. of wells positive for A only + no of wells positive for G only}}$$

[c] "Euploid" was assigned when the Pr was below the unclassifiable region; "T21" was assigned when the Pr was abov e the unclassifiable region. T21, trisomy 21

FIG. 15A

panel #	no. of wells	no. wells positive for individual alleles			
		A only	G only	GA	all negative
S1	765	14	15	0	736
S2	765	7	10	0	748
S3	765	13	9	1	744
S4	765	7	8	0	750
S5	765	12	8	0	745
S6	765	8	14	0	743
S7	765	13	9	0	743
S8	765	9	8	0	748
S9	765	14	15	0	736
S10	765	9	12	0	744
S11	765	8	17	0	740
S12	765	11	9	0	745

FIG. 15B

sample	geno-type	panels[a]	no. of wells positive for individual alleles					SPRT result		
			no of wells	A only	G only	GA	all negative	P_r	unclassifiable region	classification
M2879P	AG	5	3825	82	103	5	3635	0.557	0.570- 0.602	Euploid
M2875P	AG	3	2295	35	40	0	2220	0.533	0.546- 0.625	Euploid
M3078P	AG	3	2295	33	33	1	2228	0.500	0.541- 0.631	Euploid
M2976P	AG	12	9180	56	46	0	9078	0.549	0.556- 0.615	Euploid
M2831P	AAG	2	1530	42	13	0	1475	0.764	0.531- 0.640	T21

[a] no. of panels needed for classification

FIG. 15C

Sample	Genotype[a]	No. of wells positive for A or G allele			m_r[b]	SPRT analysis		
		A only	G only	Both		P_r[c]	Unclassifiable region[d]	Classification[e]
N3505	AG	157	131	48	0.27	0.545	0.59-0.61	Euploid
N3530	AG	174	132	52	0.28	0.569	0.59-0.61	Euploid
N3531	AG	146	129	28	0.23	0.531	0.58-0.60	Euploid
N3533	AG	149	135	46	0.27	0.525	0.59-0.61	Euploid
N2772	AG	159	149	79	0.35	0.516	0.60-0.61	Euploid
N3506	AG	154	145	66	0.32	0.515	0.60-0.61	Euploid
N3555	AG	94	73	7	0.11	0.563	0.58-0.61	Euploid
N3560	AG	101	109	28	0.18	0.519	0.58-0.61	Euploid
N3561	AG	91	70	17	0.12	0.565	0.58-0.61	Euploid
N828	AAG	153	89	22	0.16	0.632	0.58-0.60	T18
CA2	AGG	12	28	0	0.016	0.700	0.51-0.66	T18
CA7	AGG	97	171	56	0.22	0.638	0.59-0.61	T18

[a] Genotypes were determined by mass spectrometry.

[b] The m_r value indicates the average number of reference template molecules per well.

[c] The P_r value indicates the proportion of informative wells positive for the overrepresented allele.

[d] The unclassifiable region varies with the m_r value.

[e] "Euploid" was assigned when P_r was below the unclassifiable region; "T18" was assigned when P_r was above the unclassifiable region.

FIG. 16A

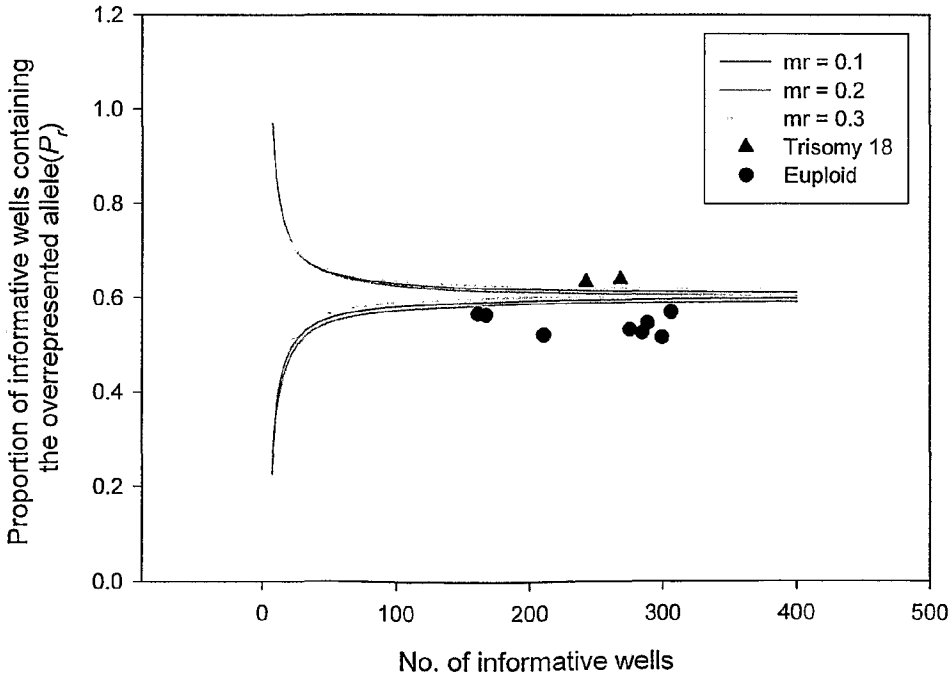


FIG. 16B

Type	Cases	Total No. of panels used	Total No. of informative wells	No. of well count			SPRT analysis			
				Both positive	Chr21 only	Chr1 only	m_r[a]	P_r[b]	Unclassifiable region	Classification[c]
T21	N0234	1	303	68	172	131	0.30	0.568	0.507 - 0.568	T21
	N0274	2	558	118	307	251	0.28	0.550	0.518 - 0.547	T21
	N0332	1	293	73	168	125	0.30	0.573	0.506 - 0.560	T21
	N0349	2	574	107	322	252	0.27	0.561	0.518 - 0.546	T21
	N0435	1	299	87	178	123	0.29	0.588	0.508 - 0.558	T21
	N0458	1	318	88	189	129	0.30	0.594	0.508 - 0.558	T21
	N0667	2	802	134	345	257	0.30	0.573	0.519 - 0.546	T21
	N0764	1	289	62	167	122	0.28	0.578	0.505 - 0.560	T21
	N0890	4	1208	272	857	549	0.31	0.545	0.528 - 0.539	T21
	N0891	1	317	103	180	137	0.38	0.568	0.510 - 0.558	T21
Normal	N0186	1	262	68	111	151	0.34	0.424	0.504 - 0.563	Euploid
	N0198	1	285	88	132	153	0.34	0.463	0.506 - 0.561	Euploid
	N0230	2	527	109	272	255	0.27	0.516	0.517 - 0.548	Euploid
	N0231	2	542	98	275	287	0.27	0.507	0.517 - 0.547	Euploid
	N0262	1	295	57	144	151	0.32	0.488	0.508 - 0.560	Euploid
	N0463	1	263	70	120	143	0.33	0.456	0.504 - 0.563	Euploid
	N0527	1	292	66	137	155	0.32	0.468	0.506 - 0.560	Euploid
	N0819	1	281	80	138	143	0.31	0.491	0.505 - 0.561	Euploid
	N0825	1	291	80	134	157	0.33	0.480	0.507 - 0.560	Euploid
	N0851	2	566	118	287	278	0.30	0.507	0.519 - 0.547	Euploid

[a] no. of reference molecules per well

[b] P_r values were calculated by the equation:

$$\frac{\text{no. of wells positive for the overrepresented allele (chr21)}}{\text{no. of wells positive for chr21 only + no of wells positive for chr1 only}}$$

[c] "Euploid" was assigned when the Pr was below the unclassifiable region; "T21" was assigned when the Pr was above the unclassifiable region.

FIG. 17

Proportion of informative wells positive for chromosome 21
0.0
0.2
0.4
0.6
0.8
1.0
1.2
0
200
400
600
800
1000
1200
1400
1600
No. of informative wells
m_r = 0.25
m_r = 0.30
m_r = 0.35
T21
Euploid

FIG. 18

Type	Cases	No. of panels used	No. of informative wells	No. of well count				SPRT analysis		
				Both positive	Chr21 only	Chr1 only	m_r[a]	P_r[b]	Unclassifiable region	Classification[c]
T21	A0274	1	380	269	249	131	0.74	0.655	0.567 - 0.585	T21
	A2008	1	348	373	233	115	1.02	0.670	0.579 - 0.596	T21
	A2178	1	384	286	239	145	0.83	0.622	0.571 - 0.588	T21
	AS363A	1	364	109	244	120	0.38	0.670	0.551 - 0.573	T21
	AS979A	1	256	492	198	88	1.32	0.734	0.590 - 0.610	T21
	AS3457A	1	376	83	258	120	0.31	0.681	0.549 - 0.572	T21
Normal	A0236	1	387	275	195	182	0.84	0.504	0.576 - 0.592	Euploid
	A0237	1	392	228	184	208	0.84	0.469	0.572 - 0.588	Euploid
	A0255	1	378	222	185	183	0.75	0.516	0.568 - 0.585	Euploid
	A0288	1	264	43	127	127	0.25	0.500	0.541 - 0.578	Euploid
	AS365A	1	370	135	192	178	0.53	0.519	0.558 - 0.578	Euploid
	AS483A	1	354	84	188	188	0.38	0.525	0.550 - 0.574	Euploid
	AS484A	1	372	271	194	178	0.88	0.522	0.573 - 0.590	Euploid
	AS780A	1	413	188	200	213	0.74	0.484	0.588 - 0.584	Euploid
	AS780A	1	316	67	159	157	0.36	0.503	0.548 - 0.575	Euploid
	AS897A	1	198	8	87	111	0.17	0.439	0.532 - 0.579	Euploid
	AS921A	1	152	8	89	83	0.13	0.454	0.523 - 0.586	Euploid
	AS980A	1	394	137	191	203	0.58	0.485	0.581 - 0.590	Euploid
	AS2282A	1	334	342	177	157	1.08	0.530	0.590 - 0.607	Euploid
	AS2425A	1	128	8	84	84	0.10	0.500	0.518 - 0.591	Euploid
	AS2434A	1	379	97	203	176	0.44	0.536	0.555 - 0.575	Euploid
	A1119	1	354	123	185	189	0.48	0.523	0.555 - 0.577	Euploid
	A1122	1	369	100	195	174	0.44	0.528	0.554 - 0.576	Euploid
	A1152	1	256	48	128	128	0.28	0.500	0.542 - 0.578	Euploid
	A1180	1	248	498	133	113	1.61	0.541	0.604 - 0.622	Euploid
	A1275	1	373	95	201	172	0.43	0.539	0.554 - 0.575	Euploid
	A1375	1	369	55	190	168	0.35	0.528	0.550 - 0.573	Euploid
	A1398	1	357	87	165	182	0.47	0.462	0.555 - 0.577	Euploid
	A1398	1	371	114	197	174	0.47	0.531	0.556 - 0.577	Euploid

[a] no. of reference molecules per well

[b] P_r values were calculated by the equation:

$$\frac{\text{no. of wells positive for the overrepresented allele (chr21)}}{\text{no. of wells positive for chr21 only + no of wells positive for chr1 only}}$$

[c] "Euploid" was assigned when the Pr was below the unclassifiable region; "T21" was assigned when the Pr was above the unclassifiable region.

FIG. 19

Type	Cases	Total No. of panels used	Total No. of informative wells	No. of well count			m_r[a]	SPRT analysis		
				Both positive	Chr18 only	Chr21 only		P_r[b]	Unclassifiable region	Classification[c]
Euploid	N0230	1	380	243	181	199	0.86	0.476	0 .573 - 0.589	E
	N0625	1	370	225	179	191	0.78	0.484	0 .569 - 0.586	E
	N0651	1	358	206	192	166	0.67	0.536	0.564 - 0.583	E
T18	N0828	1	543	256	329	214	0.95	0.606	0.579 - 0.590	T18
	N0830	1	346	353	240	106	0.92	0.694	0.574 - 0.592	T18
	N0936	1	372	285	209	163	0.88	0.562	0.573 - 0.590	E
	N1678	1	343	115	221	122	0.37	0.644	0.551 - 0.574	T18
	N1701	1	303	94	191	112	0.31	0.630	0.546 - 0.575	T18

[a] no. of reference molecules per well

[b] P_r values were calcualated by the equation:

$$\frac{\text{no. of wells positive for the overrepresented allele (chr18)}}{\text{no. of wells positive for chr21 only + no of wells positive for chr18 only}}$$

[c] "Euploid" was assigned when the Pr was below the unclassifiable region; "T18" was assigned when the Pr was above the unclassifiable region.

FIG. 20

Type	Cases	Plate No.	Assay	Total No. of informative wells	No. of well count			m_r[a]	SPRT analysis		
					Both positive	Chr 21 only	Ref only		P_r[b]	Unclassifiable region	Classification[c]
Normal	N0196	1	chr21/chr1	148	78	71	77	0.52	0.480	0.489 - 0.584	E
		1	chr21/chr18	152	63	79	73	0.44	0.520	0.487 - 0.584	U
			SUM	300	141	150	150	0.48	0.500	0.512 - 0.560	E
	N0198	1	chr21/chr1	162	73	70	92	0.56	0.432	0.490 - 0.581	E
		1	chr21/chr18	183	68	93	90	0.53	0.508	0.498 - 0.575	U
			SUM	345	141	163	182	0.55	0.472	0.517 - 0.557	E
	N0230	1	chr21/chr1	162	58	81	81	0.45	0.500	0.490 - 0.581	U
		1	chr21/chr18	165	45	81	84	0.41	0.491	0.489 - 0.580	U
			SUM	327	103	162	165	0.43	0.495	0.512 - 0.558	E
	N0231	1	chr21/chr1	173	53	81	92	0.47	0.468	0.494 - 0.578	E
		1	chr21/chr18	166	51	85	81	0.42	0.512	0.490 - 0.580	U
			SUM	339	104	166	173	0.45	0.490	0.514 - 0.557	E
	N0262	1	chr21/chr1	186	97	101	85	0.64	0.543	0.503 - 0.575	U
		1	chr21/chr18	176	111	91	85	0.71	0.517	0.504 - 0.557	U
			SUM	362	208	192	170	0.68	0.530	0.522 - 0.557	U
		2	chr21/chr1	174	103	85	89	0.69	0.489	0.503 - 0.577	E
			SUM	536	311	277	259	0.68	0.517	0 .523 - 0.552	E
T21	N0234	1	chr21/chr1	193	68	108	85	0.51	0.560	0.500 - 0.573	U
		1	chr21/chr18	178	65	92	86	0.50	0.517	0.496 - 0.576	U
			SUM	371	133	200	171	0.50	0.539	0.517 - 0.556	U
		2	chr21/chr1	178	73	92	86	0.53	0.517	0.498 - 0.576	U
			SUM	549	206	292	257	0.51	0.532	0 .524 - 0.549	U
		2	chr21/chr18	178	62	95	83	0.47	0.534	0.495 - 0.576	U
			SUM	727	268	387	340	0.50	0.532	0 .527 - 0.546	U
		3	chr21/chr1	184	108	103	81	0.68	0.560	0.504 - 0.575	U
			SUM	911	376	490	421	0.54	0.538	0.529 - 0.545	U
		3	chr21/chr18	198	102	120	78	0.63	0.606	0.505 - 0.572	T21
			SUM	1109	478	610	499	0.55	0.550	0.531 - 0.544	T21
	N0332	1	chr21/chr1	173	109	103	70	0.63	0.595	0.500 - 0.577	T21
		1	chr21/chr18	191	82	110	81	0.55	0.576	0.501 - 0.574	T21
			SUM	364	191	213	151	0.59	0.585	0.519 - 0.557	T21
	N0349	1	chr21/chr1	158	62	83	75	0.44	0.525	0.489 - 0.582	U
		1	chr21/chr18	166	46	87	79	0.39	0.524	0.489 - 0.580	U
			SUM	324	108	170	154	0.42	0.525	0.512 - 0.558	U
		2	chr21/chr1	149	66	86	63	0.41	0.577	0.485 - 0.585	U
			SUM	473	174	256	217	0.41	0.541	0.519 - 0.551	U
		2	chr21/chr18	168	43	100	68	0.34	0.595	0.487 - 0.580	T21
			SUM	641	217	356	285	0.40	0.555	0.523 - 0.546	T21
	N0764	1	chr21/chr1	182	82	106	76	0.53	0.582	0.498 - 0.575	T21
		1	chr21/chr18	178	83	97	81	0.56	0.545	0.498 - 0.576	U
			SUM	360	165	203	157	0.54	0.564	0.518 - 0.556	T21
	N0890	1	chr21/chr1	180	103	108	72	0.61	0.600	0.501 - 0.576	T21
		1	chr21/chr18	188	68	104	84	0.50	0.553	0.499 - 0.574	U
			SUM	368	171	212	156	0.55	0.576	0.518 - 0.556	T21

[a] no. of reference molecules per well

[b] P_r values were calcualated by the equation:

$$\frac{\text{no. of wells positive for the overrepresented allele (chr21)}}{\text{no. of wells positive for chr21 only + no of wells positive for Ref only}}$$

[c] "Euploid" was assigned when the Pr was below the unclassifiable region; "T21" was assigned when the Pr was above the unclassifiable region.

FIG. 21

			Poisson corrected total no. Chromosome 21	Reference chr	P_r	Classification
Euploid	Case 1	Assay 1	37	33	0.52345751	**Unclass**
		Assay 2	31	42	0.42506276	**Unclass**
		Assay 3	29	37	0.44198785	**Unclass**
		Assay 4	39	26	0.6008691	**Unclass**
		Plate 1 subtotal	136	138	0.495821	**Unclass**
		Assay 1	32	25	0.56605316	**Unclass**
		Assay 2	27	24	0.53165424	**Unclass**
		Assay 3	10	17	0.3681535	**Unclass**
		Assay 4	24	18	0.5626764	**Unclass**
		Plate 2 subtotal	93	84	0.52474	**Unclass**
		Total	229	223	0.507182	**Euploid**
	Case 2	Assay 1	45	35	0.56321585	**Unclass**
		Assay 2	31	45	0.41245525	**Unclass**
		Assay 3	32	22	0.59954763	**Unclass**
		Assay 4	38	35	0.5228175	**Unclass**
		Plate 1 subtotal	146	135	0.519153	**Unclass**
		Assay 1	28	23	0.55272135	**Unclass**
		Assay 2	23	22	0.51196923	**Unclass**
		Assay 3	23	15	0.59688187	**Unclass**
		Assay 4	29	26	0.5293349	**Unclass**
		Plate 2 subtotal	102	85	0.545204	**Unclass**
		Plate 1+2 subtotal	248	221	0.529583	**Unclass**
		Assay 1	17	25	0.4122149	**Unclass**
		Assay 2	25	16	0. 6027299	**Unclass**
		Assay 3	13	26	0.3382802	**Unclass**
		Assay 4	27	24	0.5316872	**Unclass**
		Plate 3 subtotal	82	91	0.475775	**Unclass**
		Plate 1- 3 subtotal	331	311	0.515069	**Unclass**
		Assay 1	15	9	0.6269193	**Unclass**
		Assay 2	14	15	0.4824155	**Unclass**
		Assay 3	18	17	0.5146313	**Unclass**
		Assay 4	21	18	0.5270080	**Unclass**
		Plate 4 subtotal	69	60	0.532264	**Unclass**
		Plate 1- 4 subtotal	399	372	0.517942	**Unclass**

FIG. 22A

	Assay 1	10	13	0.4337957	**Unclass**
	Assay 2	11	14	0.4390082	**Unclass**
	Assay 3	7	16	0.3018022	**Unclass**
	Assay 4	19	10	0.6579085	**Unclass**
	Plate 5 subtotal	48	54	0.469951	**Unclass**
	Total	447	426	0.5123444	**Euploid**
T21					
Case 3	Assay 1	50	33	0.60012253	**Unclass**
	Assay 2	56	47	0.54458055	**Unclass**
	Assay 3	46	28	0.6196626	**Unclass**
	Assay 4	56	36	0.6108731	**Unclass**
	Total	208	144	0.59079	**T21**
Case 4	Assay 1	56	36	0.61085513	**Unclass**
	Assay 2	38	32	0.53906912	**Unclass**
	Assay 3	40	24	0.62788068	**Unclass**
	Assay 4	47	27	0.634511	**Unclass**
	Total	180	119	0.603476	**T21**

FIG. 22B

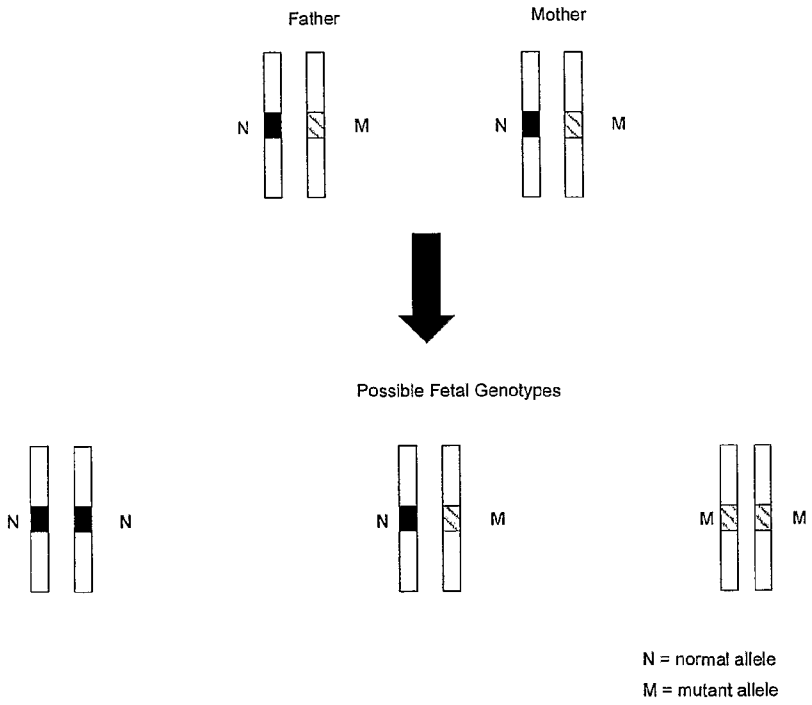

Father
Mother
N
M
N
M
Possible Fetal Genotypes
N
N
N
M
M
M
N = normal allele
M = mutant allele

FIG. 23

Cases	Samples	Expected %	No of panels	*ZFY*	*ZFX*	ALL POS	NET *ZFY*	NET *ZFX*	Ref allele	m_r	SPRT[a]
1	XY + XX	50%	1	239	465	155	84	310	Y	0.37	XX
	(XX %)	25%	1	328	447	205	123	242	Y	0.56	XX
2	$XY_1 + XY_2$	50%	1	414	405	220	194	185	X	0.85	XY
	(XY_2 %)	25%	1	432	426	245	187	181	X	0.81	XY

FIG. 24A

Cases	No of panels	*ZFY*	*ZFX*	ALL POS	NET *ZFY*	NET *ZFX*	Ref allele	m_r	SPRT
1	1	81	171	18	63	153	Y	0.11	XX
2	1	103	175	23	80	152	Y	0.14	XX
3	1	110	175	28	82	147	Y	0.16	XX
4	1	111	156	24	87	132	Y	0.16	XX
5	2	236	322	50	186	272	Y	0.17	XX
6	1	116	184	29	87	155	Y	0.16	XX
7	1	94	169	19	75	150	Y	0.13	XX
8	1	91	155	24	67	131	Y	0.13	XX
9	3	333	447	74	259	373	Y	0.16	XX
10	1	110	154	24	86	130	Y	0.16	XX
11	1	83	170	26	57	144	Y	0.11	XX
12	1	85	159	16	69	143	Y	0.12	XX

FIG. 24B

Cases	No of panels	Maternal genotype[a]	Fetal genotype[a]	Expected fetal %	A	G	ALL POS	NET A	NET G	Actual fetal %[b]	Ref allele	m_t	SPRT[d]
1	1	AG	GG	15%	264	318	116	148	202	9.6	A	0.42	GG
2	1	AG	GG	15%	208	276	81	127	195	14.3	A	0.32	GG
3	1	AG	AG	15%	248	251	87	161	164	15.4	A	0.39	AG
4	1	AG	GG	15%	158	240	51	107	189	13.7	A	0.23	GG
5	1	AG	AG	15%	226	245	70	156	175	16.8	A	0.35	AG
6	1	AG	AG	15%	319	319	130	189	189	18.1	A	0.54	AG
7	1	AG	AG	15%	254	279	94	160	185	19.8	A	0.40	AG
8	1	AG	GG	15%	296	363	144	152	219	15.7	A	0.49	GG
9	1	AG	AG	15%	279	285	100	179	185	13.0	A	0.45	AG
10	1	AG	AA	15%	337	278	120	217	158	NA[c]	G	0.45	AA
1	1	AG	GG	20%	248	295	96	152	199	13.0	A	0.39	GG
2	1	AG	GG	20%	191	289	79	112	210	19.5	A	0.29	GG
3	1	AG	AG	20%	248	218	75	173	143	18.2	G	0.34	U
4	1	AG	GG	20%	158	257	59	99	198	25.9	A	0.23	GG
5	1	AG	AG	20%	231	215	71	160	144	20.3	G	0.33	AG
6	1	AG	AG	20%	281	283	98	183	185	26.3	A	0.46	AG
7	1	AG	AG	20%	256	231	88	168	143	22.7	G	0.36	AG
8	1	AG	GG	20%	246	311	104	142	207	23.0	A	0.39	GG
9	1	AG	AG	20%	267	249	83	184	166	21.9	G	0.39	AG
10	1	AG	AA	20%	303	203	77	226	126	NA[c]	G	0.31	AA
1	1	AG	GG	25%	244	304	99	145	205	15.7	A	0.38	GG
2	1	AG	GG	25%	209	314	88	121	226	24.9	A	0.32	GG
3	1	AG	AG	25%	204	212	55	149	157	28.0	A	0.31	AG
4	1	AG	GG	25%	185	290	76	109	214	25.9	A	0.28	GG
5	1	AG	AG	25%	233	243	77	156	166	24.5	A	0.36	AG
6	1	AG	AG	25%	292	305	110	182	195	27.3	A	0.48	AG
7	1	AG	AG	25%	261	275	95	166	180	22.5	A	0.42	AG
8	1	AG	GG	25%	236	327	96	140	231	26.7	A	0.37	GG
9	1	AG	AG	25%	271	280	104	167	176	25.2	A	0.44	AG
10	1	AG	AA	25%	243	154	46	197	108	NA[c]	G	0.22	AA
1	1	AG	GG	50%	136	293	57	79	236	33.6	A	0.20	GG
2	1	AG	GG	50%	163	358	78	85	280	45.8	A	0.24	GG
3	1	AG	AG	50%	200	208	53	147	155	55.7	A	0.30	AG
4	1	AG	GG	50%	137	356	66	71	290	53.0	A	0.18	GG
5	1	AG	AG	50%	247	246	91	156	155	57.7	G	0.39	AG
6	1	AG	AG	50%	218	245	68	150	177	53.8	A	0.34	AG
7	1	AG	AG	50%	255	282	103	152	179	51.3	A	0.41	AG
8	1	AG	GG	50%	158	398	83	75	315	57.6	A	0.23	GG
9	1	AG	AG	50%	235	225	74	161	151	55.7	G	0.35	AG
10	1	AG	AA	50%	317	106	35	282	71	NA[c]	G	0.15	AA

[a] A = Mutated allele, G = Wildtype allele.
[b] The actual fetal DNA % is determined by the digital *ZFY/X* assay.
[c] The fetus is female and digital *ZFY/X* assay cannot be applied.
[d] SPRT was done with the fetal % determined by the digital *ZFY/X* assay except case 10; U, unclassified.

FIG. 25

Case	Maternal genotype[a]	Fetal genotype[a]	No of panels	M	N	ALL POS	NET M	NET N	1[b] Expected fetal %	1[b] SPRT	2[c] Fetal % by digital *ZFY/X*	2[c] SPRT	Ref Allele	m_r
1	MN	MM	3	846	763	273	573	490	5%	MM	7.8%	MM	N	0.40
	MN	MM	3	863	748	289	574	459	10%	MM	9.6%	MM	N	0.39
	MN	MM	1	323	219	79	244	140	25%	MM	24.1%	MM	N	0.34
	MN	MM	1	418	160	92	326	68	50%	MM	57.2%	MM	N	0.23
2	MN	MN	9	2559	2476	887	1672	1589	5%	U	3.4%	U	N	0.45
	MN	MN	3	895	907	360	535	547	10%	MN	9.7%	MN	M	0.49
	MN	MN	1	277	263	97	180	166	25%	MN	24.2%	MN	N	0.42
	MN	MN	1	273	295	101	172	194	50%	MN	58.0%	MN	M	0.44

[a] M = with CTTT deletion, N = without CTTT deletion.
[b] SPRT was done with the expected %.
[c] SPRT was done with the fetal % determined by the digital *ZFY/X* assay; U, unclassified

FIG. 26A

Cases	No of panels	Maternal genotype[a]	Fetal genotype[a]	Expected fetal %	M	N	ALL POS	NET M	NET N	Actual fetal %[b]	Ref allele	m_r	SPRT[c]
1	1	MN	MN	20%	203	200	58	145	142	29.0	N	0.30	MN
2	1	MN	NN	20%	149	204	39	110	165	29.7	M	0.22	NN
3	1	MN	NN	20%	128	212	31	97	181	30.7	M	0.18	NN
4	1	MN	MN	20%	173	196	45	128	151	27.0	M	0.26	MN
5	1	MN	NN	20%	147	230	47	100	183	24.3	M	0.21	NN
6	1	MN	MN	20%	199	206	45	154	161	19.2	M	0.30	MN
7	1	MN	NN	20%	165	237	58	107	179	18.7	M	0.24	NN
8	1	MN	MN	20%	203	180	43	160	137	23.4	N	0.27	MN
9	1	MN	NN	20%	173	241	61	112	180	16.9	M	0.26	NN
10	1	MN	MN	20%	174	190	52	122	138	16.8	M	0.26	U
11	1	MN	NN	20%	167	242	49	118	193	17.7	M	0.25	NN
12	1	MN	MN	20%	213	199	47	166	152	21.6	N	0.30	MN

[a] M = with CTTT deletion, N = without CTTT deletion.
[b] The actual fetal DNA % is determined by the digital *ZFXY* assay.
[c] SPRT was done with the fetal % determined by the digital *ZFY/X* assay; U, unclassified.

FIG. 26B

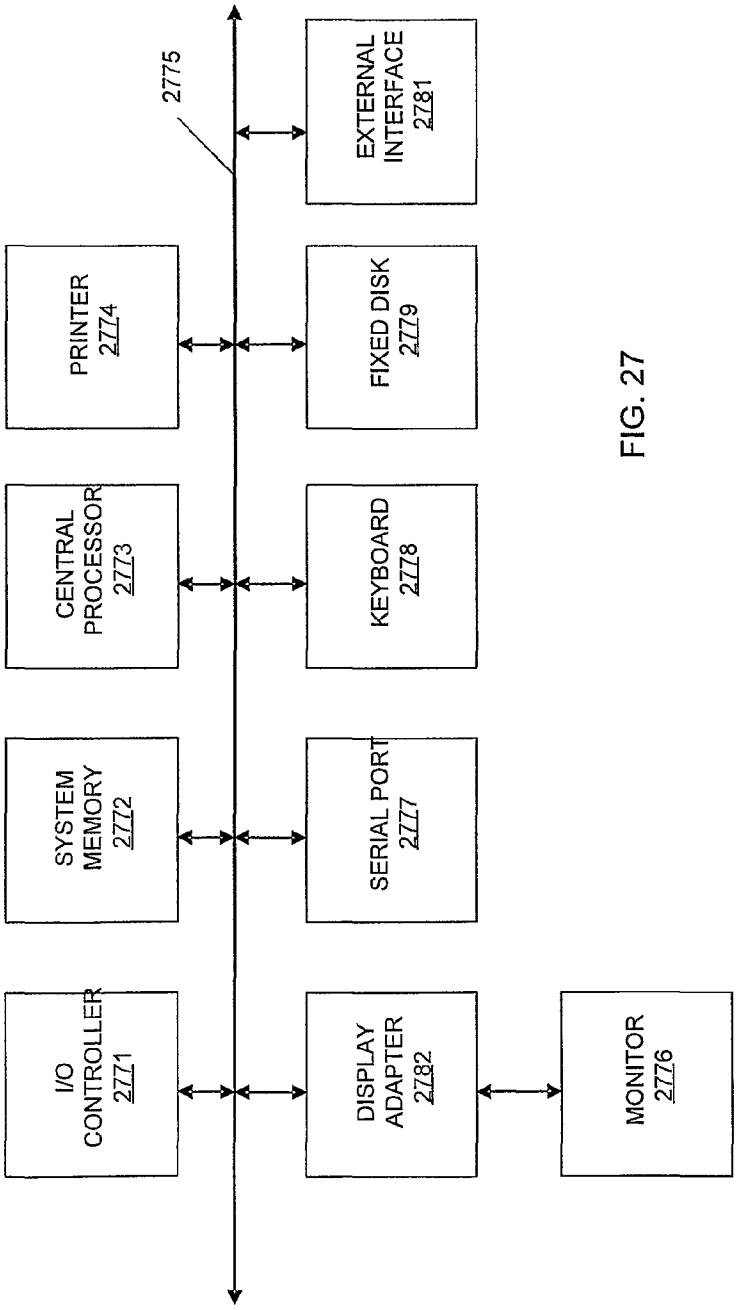

2775
EXTERNAL INTERFACE 2781
PRINTER 2774
FIXED DISK 2779
CENTRAL PROCESSOR 2773
KEYBOARD 2778
SYSTEM MEMORY 2772
SERIAL PORT 2777
I/O CONTROLLER 2771
DISPLAY ADAPTER 2782
MONITOR 2776

FIG. 27

INTERNATIONAL SEARCH REPORT

International application No
PCT/GB2008/002524

A. CLASSIFICATION OF SUBJECT MATTER
INV. G06F19/00 C12Q1/68

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)
G06F C12Q

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

Electronic data base consulted during the international search (name of data base and, where practical, search terms used)
EPO-Internal, BIOSIS, EMBASE, COMPENDEX, INSPEC, WPI Data

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	ZHOU W ET AL: "Counting alleles to predict recurrence of early-stage colorectal cancers" LANCET THE, LANCET LIMITED. LONDON, GB, vol. 359, no. 9302, 19 January 2002 (2002-01-19), pages 219-225, XP004791874 ISSN: 0140-6736 cited in the application abstract page 220, right-hand column, line 38 - page 221, right-hand column, line 10 figures 1,2	1-12, 14-18, 25,26, 30,31

-/--

[X] Further documents are listed in the continuation of Box C. [ ] See patent family annex.

* Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance
"E" earlier document 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
24 October 2008	30/10/2008

Name and mailing address of the ISA/	Authorized officer
European Patent Office, P.B. 5818 Patentlaan 2 NL – 2280 HV Rijswijk Tel. (+31-70) 340-2040, Fax: (+31-70) 340-3016	Kürten, Ivayla

Form PCT/ISA/210 (second sheet) (April 2005)

C(Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	SHIH I-M ET AL: "Evidence that genetic instability occurs at an early stage of colorectal tumorigenesis" CANCER RESEARCH, vol. 61, February 2002 (2002-02), pages 818-822, XP007905911 page 819, right-hand column, line 13 - page 820, right-hand column, line 21 figure 2	1-12, 14-18, 25,26, 30,31
X	TONG YU K ET AL: "Noninvasive prenatal detection of fetal trisomy 18 by epigenetic allelic ratio analysis in maternal plasma: Theoretical and empirical considerations" CLINICAL CHEMISTRY, AMERICAN ASSOCIATION FOR CLINICAL CHEMISTRY, WASHINGTON, DC, vol. 52, no. 12, 13 October 2006 (2006-10-13), pages 2194-2202, XP002470084 ISSN: 0009-9147 cited in the application	1-12, 14-18,31
A	page 2197, left-hand column, line 14 - page 2199, right-hand column, line 32 page 2201 figure 2	25,26,30
P,X	LO YMD ET AL: "Digital PCR for the molecular detection of fetal chromosomal aneuploidy." PNAS, vol. 104, no. 32, 7 August 2007 (2007-08-07), pages 13116-13121, XP007905909 page 13117, left-hand column - page 13118, right-hand column	1-12, 14-18, 25,26, 30,31
A	LO YMD ET AL: "Plasma placental RNA allelic ratio permits noninvasive prenatal chromosomal aneuploidy detection" NATURE MEDICINE, vol. 13, no. 2, February 2007 (2007-02), pages 218-223, XP007905910 cited in the application page 219, left-hand column, line 5 - line 25 page 219, right-hand column, line 26 - page 221, left-hand column page 222, right-hand column, line 13 - line 24	1-12, 14-18, 25,26, 30,31

-/--

C(Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	EL KAROUI N ET AL: "Getting more from digital SNP data" STATISTICS IN MEDICINE, vol. 25, 2006, pages 3124-3133, XP007905913 cited in the application page 3125, line 25 - page 3128, line 19	1-12, 14-18, 25,26, 30,31
A	LO YMD AND CHIU RWK: "Prenatal diagnosis: progress through plasma nucleic acids." NATURE REVIEWS GENETICS, vol. 8, January 2007 (2007-01), pages 71-77, XP007905874 cited in the application the whole document	1-12, 14-18, 25,26, 30,31

INTERNATIONAL SEARCH REPORT

International application No.
PCT/GB2008/002524

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.: 13,19-24,27-29
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:
see FURTHER INFORMATION sheet PCT/ISA/210

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

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

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

see additional sheet

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

2. ☒ As all searchable claims could be searched without effort justifying an 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 reportcovers 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.

FURTHER INFORMATION CONTINUED FROM PCT/ISA/ 210

Continuation of Box II.2

Claims Nos.: 13,19-24,27-29

No opinion is given for claims 13, 19 to 24 and 27 to 29 because the subject-matter of said claims is not clearly defined contrary to the requirements of Article 6 PCT. The reasons are as follows:

According to claim 13 "the cutoff value is based on a measurement of one of the first portions or a measurement of one of the second portions". This vague expression does not enable the skilled reader to determine how the cutoff value is calculated nor which quantities are used in the calculation of said cutoff value. The subject-matter of claim 13 is unclear to such an extent that no meaningful search could be carried out.

According to claim 19, the "calculating of the first cutoff value includes at least one of: sequential probability ratio testing, false discovery rates, confidence intervals, and receiver operating characteristic curves." These listed methods allow for hypothesis testing (i.e. sequence imbalance vs sequence balance) based on given cutoff values. It is unclear how these methods are used for calculating the cutoff value itself.

Claims 20 to 24 define a proportion P1 and a set of probability values and state that the method includes "calculating the first cutoff value from the first proportion P1" (claim 20). This information, however, is insufficient to define said cutoff value. Moreover, the calculation of P1 itself is unclear: according to claim 21, P1 is calculated "using the first probability and the second probability". It is unclear how exactly these probabilities are used in the calculation of P1.

In claim 22, it is unclear what is being addressed by "an expected ratio relative to the non-reference nucleic acid sequence". Consequently, the calculation of the "first probability" remains unclear as well.

Claim 24 defines yet another probability, "third probability" but does not specify the purpose of this probability. It is unclear why this probability is defined and how it is used in the determination of the cutoff value. Also, according to claim 24 "calculating the second probability ... includes assuming that first probability and second probability are independent". It is unclear how the independence of the first and the second probabilities is incorporated in the calculation of the second probability.

The wording of claims 19 to 24 does not enable the skilled person to determine how the cutoff value is calculated. Consequently, no search regarding the definition of the cutoff value could be carried out. Although the description is more specific regarding the calculation of the cutoff values, the relevant passages are spread across pages 22 to 26 and could not be interpreted unambiguously.

Claim 27 also attempts to define the first cutoff value by "determining

FURTHER INFORMATION CONTINUED FROM PCT/ISA/ 210

a first average concentration ...; and multiplying the first average concentration by a factor derived from the first percentage to obtain a second average concentration ...". The claim only specifies that the factor is derived from the first percentage, but does not specify how exactly said factor is calculated. Furthermore, the result of the claim is a second average concentration. The claim does not specify how this second average concentration is used in the calculation of the cutoff value.

Claims 28 and 29 further specify the average concentration but do not enlighten the procedure of calculating the cutoff value.

Consequently, the method of calculating said cutoff value is unclear to such an extent that could not be meaningfully searched.

The applicant's attention is drawn to the fact that claims relating to inventions in respect of which no international search report has been established need not be the subject of an international preliminary examination (Rule 66.1(e) PCT). The applicant is advised that the EPO policy when acting as an International Preliminary Examining Authority is normally not to carry out a preliminary examination on matter which has not been searched. This is the case irrespective of whether or not the claims are amended following receipt of the search report or during any Chapter II procedure. If the application proceeds into the regional phase before the EPO, the applicant is reminded that a search may be carried out during examination before the EPO (see EPO Guideline C-VI, 8.2), should the problems which led to the Article 17(2)PCT declaration be overcome.

FURTHER INFORMATION CONTINUED FROM PCT/ISA/ 210

This International Searching Authority found multiple (groups of) inventions in this international application, as follows:

1. claims: 1-12,14-18,31

A method for determining whether a nucleic acid sequence imbalance in a sample exists whereby the average concentration of a reference nucleic acid in a reaction is taken into account.

---

2. claims: 25-26,30

A method for determining whether a nucleic acid sequence imbalance in a sample exists whereby the sample consists of two or more types of cells and the percentage of the nucleic acid sequence from the first type of cells is taken into account.

Electronic Acknowledgement Receipt

EFS ID:	40964371
Application Number:	16995607
International Application Number:	
Confirmation Number:	1256
Title of Invention:	DETERMINATION OF BASE MODIFICATIONS OF NUCLEIC ACIDS
First Named Inventor/Applicant Name:	Yuk-Ming Dennis Lo
Customer Number:	20350
Filer:	David Dah-Wei Hsu/Heather Orosco
Filer Authorized By:	David Dah-Wei Hsu
Attorney Docket Number:	080015-028410US-1199353
Receipt Date:	27-OCT-2020
Filing Date:	17-AUG-2020
Time Stamp:	20:10:28
Application Type:	Utility under 35 USC 111(a)

Payment information:

Submitted with Payment	no

File Listing:

Document Number	Document Description	File Name	File Size(Bytes)/ Message Digest	Multi Part /.zip	Pages (if appl.)
1		1199353_Initial_IDS.pdf	174473 e074030babddbf73228bb8cbf90f22ea09bc541e	yes	5

	Multipart Description/PDF files in .zip description				
	Document Description		Start	End	
	Information Disclosure Statement (IDS) Form (SB08)		3	5	
	Transmittal Letter		1	2	
Warnings:					
Information:					
2	Foreign Reference	REF_English_abstract_and_KR101896167B1.pdf	5312542 27c6a31d26d2fa99872437320b9e9036c6ea1851	no	98
Warnings:					
Information:					
3	Foreign Reference	REF_WO2009013492A1.pdf	6282677 140035e7548efc64261c1c5ed48382258d0c5477	no	115
Warnings:					
Information:					
4	Non Patent Literature	REF_Flusberg_2010_Direct_detection_of_DNA_methylation.pdf	398818 2c6d8bcc1c160f415c6857f791f5787293f0e389	no	7
Warnings:					
Information:					
5	Non Patent Literature	REF_Liu_2019_Detection_of_DNA_base_modifications.pdf	1308342 ce4a907595d1e9ebf89dfea296053af4fbd1819c	no	11
Warnings:					
Information:					
6	Non Patent Literature	REF_Ni_2019_DeepSignal__detecting_DNA_methylation.pdf	910480 b6809a4ba09fe98a961ac090a2b39fd8d40c76af	no	10
Warnings:					
Information:					
7	Non Patent Literature	REF_Suzuki_2016_Agln__measuring_the_landscape_of_CpG.pdf	678850 1c1cd22c2a23060488756ca44a5b8cf6ca5712f0	no	9
Warnings:					

Information:					
8	Non Patent Literature	REF_Blow_2016_The_Epigenomic_Landscape_of_Prokaryotes.pdf	2200692 51764d78716d45ee7e10ecc59986ebc0f29bbdda	no	28
Warnings:					
Information:					
9	Non Patent Literature	REF_Clark_2013_Enhanced_5-methylcytosine_detection_in_single-molecule.pdf	1031167 7a6dbcfa901d1ea62685b527b319527cec0a9af2	no	10
Warnings:					
Information:					
10	Non Patent Literature	REF_Clark_2011_Characterization_of_DNA_methyltransferase.pdf	1858315 243121c513bf9b402764d49317d6491186d228ea	no	12
Warnings:					
Information:					
11	Non Patent Literature	REF_Eid_2009_Real-Time_DNA_Sequencing_from_Single.pdf	930040 dea0422fae46641f562afa44ae197b187eea5609	no	7
Warnings:					
Information:					
12	Non Patent Literature	REF_Feng_2013_Detecting_DNA_Modifications_from_SMRT_Sequencing.pdf	991408 b7695d97981d28afeb463584544ffdc393351d5e	no	10
Warnings:					
Information:					
13	Non Patent Literature	REF_Liu_2019_Bisulfite-free_direct_detection.pdf	872165 af923313ce47bc297bf818dca55fd75aded58070	no	11
Warnings:					
Information:					
14	Non Patent Literature	REF_Schadt_2013_Modeling_kinetic_rate_variation_in_third.pdf	1537212 af4baf607e4989d7720f87f50bb1cfb11a58ada6	no	13
Warnings:					
Information:					

15	Non Patent Literature	REF_Feng_2014_qDNAmod_a_statistical_model-based_tool.pdf	1397844 58d06695c4c775a667e315095b4a641f23bea1bc	no	12
Warnings:					
Information:					
16	Non Patent Literature	REF_Suzuki_2017_Observing_comprehensive_DNA_methylomes_Part_1.pdf	15783425 8bb1375e8effc5dd19e428d557254390033d0e92	no	52
Warnings:					
Information:					
17	Non Patent Literature	REF_Suzuki_2017_Observing_comprehensive_DNA_methylomes_Part_2.pdf	17721789 74fdd56376a4e4d3e43dbbb84aa813352980a56e	no	66
Warnings:					
Information:					
			Total Files Size (in bytes):	59390239	

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 a 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.

DocCode – SEQ.TXT

SCORE Placeholder Sheet for IFW Content

Application Number: 16995607 Document Date: 10/27/2020

The presence of this form in the IFW record indicates that the following document type was received in electronic format on the date identified above. This content is stored in the SCORE database.

Since this was an electronic submission, there is no physical artifact folder, no artifact folder is recorded in PALM, and no paper documents or physical media exist. The TIFF images in the IFW record were created from the original documents that are stored in SCORE.

- Sequence Listing

At the time of document entry (noted above):

- USPTO employees may access SCORE content via DAV or via the SCORE web page.
- External customers may access SCORE content via PAIR using the Supplemental Content tab.

Form Revision Date: March 1, 2019

I hereby certify that this correspondence is being filed via EFS-Web with the United States Patent and Trademark Office on October 27, 2020.

KILPATRICK TOWNSEND & STOCKTON LLP

By: /Heather A. Orosco/
Heather A. Orosco

PATENT
Attorney Docket No.: 080015-028410US-1199353
Client Ref. No.: 19/MED/901 US Reg

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

In re Application of:

Yuk-Ming Dennis Lo et al.

Application No.: 16/995,607

Filed: August 17, 2020

For: DETERMINATION OF BASE MODIFICATIONS OF NUCLEIC ACIDS

Customer No.: 20350

Confirmation No.: 1256

Examiner: Not Yet Known

Technology Center/Art Unit: Not Yet Known

COMMUNICATION UNDER 37 C.F.R. §§ 1.821-1.825 and PRELIMINARY AMENDMENT

Mail Stop Missing Parts
Commissioner for Patents
P.O. Box 1450
Alexandria, VA 22313-1450

Commissioner:

Prior to examination of the above-referenced application, please enter the following amendments and consider the following remarks:

Amendments to the Specification begin on page 2 of this paper.

Remarks/Arguments begin on page 4 of this paper.

UNITED STATES PATENT AND TRADEMARK OFFICE

UNITED STATES DEPARTMENT OF COMMERCE
United States Patent and Trademark Office
Address: COMMISSIONER FOR PATENTS
P.O. Box 1450
Alexandria, Virginia 22313-1450
www.uspto.gov

APPLICATION NUMBER	FILING OR 371(C) DATE	FIRST NAMED APPLICANT	ATTY. DOCKET NO./TITLE
16/995,607	08/17/2020	Yuk-Ming Dennis Lo	080015-028410US-1199353

CONFIRMATION NO. 1256

20350
KILPATRICK TOWNSEND & STOCKTON LLP
Mailstop: IP Docketing - 22
1100 Peachtree Street
Suite 2800
Atlanta, GA 30309

FORMALITIES LETTER

OC000000120167457

Date Mailed: 09/25/2020

NOTICE TO COMPLY WITH REQUIREMENTS FOR PATENT APPLICATIONS CONTAINING NUCLEOTIDE AND/OR AMINO ACID SEQUENCE DISCLOSURES

Filing Date Granted

Applicant is given **TWO MONTHS FROM THE DATE OF THIS NOTICE** within which to comply with the sequence rules, 37 CFR §§ 1.821-1.825. Failure to comply with these requirements will result in ABANDONMENT of the application under 37 CFR §1.821(g). Extension of time may be obtained by filing a petition accompanied by the extension fee under the provisions of 37 CFR §1.136.

Items 1) and 2) provide general guidance related to requirements for sequence disclosures.

1) 37 CFR 1.821(c) requires that patent applications which contain disclosures of nucleotide and/or amino acid sequences that fall within the definitions of 37 CFR 1.821(a) must contain a "Sequence Listing", *as a separate part of the disclosure*, which presents the nucleotide and/or amino acid sequences and associated information using the symbols and format in accordance with the requirements of 37 CFR 1.821-1.825. This "Sequence Listing" part of the disclosure may be submitted:
   a) via EFS-Web (see Section I.1 of the Legal Framework for EFS-Web (https://www.uspto.gov/patents-application-process/filing-online/legal-framework-efs-web), hereinafter "Legal Framework") as an ASCII text file, together with an incorporation-by-reference of the material in the ASCII text file in a separate paragraph of the specification identifying:
      i) the name of the ASCII text file;
      ii) the date of creation; and
      iii) the size of the ASCII text file in bytes;
   b) on compact disc(s) in duplicate according to 37 CFR 1.52(e)(1)(ii) and (4), labeled according to 37 CFR 1.52(e)(6), together with a statement that the duplicate compact discs are identical and an incorporation-by-reference of the material in the ASCII text file according to 37 CFR 1.52(e)(5) in a separate paragraph of the specification identifying:
      i) the name of the ASCII text file;
      ii) the date of creation; and
      iii) the size of the ASCII text file in bytes;
   c) via EFS-Web as a PDF (not recommended); or
   d) on paper.
2) 37 CFR 1.821(e) requires that a *copy* of the "Sequence Listing" must also be submitted in computer readable form (CRF) in accordance with the requirements of 37 CFR 1.824.
   a) If a "Sequence Listing" ASCII text file submitted via EFS-Web on the application filing date complies with the requirements of 37 CFR 1.824(a)(2)-(6) and (b), and applicant has not filed a "Sequence Listing" as

a PDF, the text file will serve as both the "Sequence Listing" required by 37 CFR 1.821(c) and the CRF required by 37 CFR 1.821(e), and the statement of identity under the "Legal Framework" is not required.

b) If the "Sequence Listing" required by 37 CFR 1.821(c) is filed via EFS-Web as a PDF, then the "Legal Framework" requires submission of a statement that the "Sequence Listing" content of the PDF copy and the ASCII text file copy submitted via EFS-Web are identical.

c) If the "Sequence Listing" required by 37 CFR 1.821(c) is filed on paper or compact disc, then 37 CFR 1.821(f) requires submission of a statement that the "Sequence Listing" content of the paper or compact disc copy and the CRF are identical.

Specific deficiencies and the required response to this notice are as follows:

- This application fails to comply with the requirements of 37 CFR 1.821-1.825 because it does not contain a "Sequence Listing" as a separate part of the disclosure or a CRF. Applicant must provide:
  - A "Sequence Listing" part of the disclosure, as described above in item 1); **together with**
  - **An amendment specifically directing its entry into the application;**
  - A statement that the "Sequence Listing" includes no new matter; and
  - A statement that indicates support for the amendment in the application, as filed, as required by 37 CFR 1.825 and the "Legal Framework".
- If the "Sequence Listing" part of the disclosure is submitted according to item 1) a) or b) above, Applicant must also provide:
  - A substitute specification in compliance with 37 CFR 1.121(b)(3) and 1.125 inserting the required incorporation-by-reference paragraph, consisting of:
    - A copy of the previously-submitted specification, with deletions shown with strikethrough or brackets and insertions shown with underlining (marked-up version);
    - A copy of the amended specification without markings (clean version); and
    - A statement that the substitute specification contains no new matter.
- If the "Sequence Listing" part of the disclosure is submitted according to item 1) b), c), or d) above, Applicant must also provide:
  - A CRF; and
  - A statement according to item 2) b) or c) above.
- If applicant desires the CRF in this application to be identical with the CRF of another application on file in the U.S. Patent and Trademark Office, such request in accordance with 37 CFR 1.821(e) may be submitted in lieu of a new CRF (Form PTO/SB/93 may be used for this purpose).

For questions regarding compliance to these requirements, please contact:

- **For Rules interpretation, call (571) 272-2510**
- **For CRF submission help, call (571) 272-2510**
- **For PatentIn software program support technical assistance, call (866) 217-9197.**
- **Send e-mail correspondence for PatentIn software program help to ebc@uspto.gov**
- **PatentIn software is available at http://www.uspto.gov/patents/resources/tools/checker/patentinrel.jsp .**

Applicant is cautioned that correction of the above items may cause the specification and drawings page count to exceed 100 pages. If the specification and drawings exceed 100 pages, applicant will need to submit the required application size fee.

Replies must be received in the USPTO within the set time period or must include a proper Certificate of Mailing or Transmission under 37 CFR 1.8 with a mailing or transmission date within the set time period. For more information and a suggested format, see Form PTO/SB/92 and MPEP 512.

Replies should be mailed to:

Mail Stop Missing Parts
Commissioner for Patents
P.O. Box 1450
Alexandria VA 22313-1450

Registered users of EFS-Web may alternatively submit their reply to this notice via EFS-Web, including a copy of this Notice and selecting the document description "Applicant response to Pre-Exam Formalities Notice". https://sportal.uspto.gov/authenticate/AuthenticateUserLocalEPF.html

For more information about EFS-Web please call the USPTO Electronic Business Center at 1-866-217-9197 or visit our website at http://www.uspto.gov/ebc.

If you are not using EFS-Web to submit your reply, you must include a copy of this notice.

Questions about the contents of this notice and the requirements it sets forth should be directed to the Office of Data Management, Application Assistance Unit, at **(571) 272-4000** or **(571) 272-4200** or **1-888-786-0101**.

/sareebuddin/

Electronic Acknowledgement Receipt

EFS ID:	40961796
Application Number:	16995607
International Application Number:	
Confirmation Number:	1256
Title of Invention:	DETERMINATION OF BASE MODIFICATIONS OF NUCLEIC ACIDS
First Named Inventor/Applicant Name:	Yuk-Ming Dennis Lo
Customer Number:	20350
Filer:	David Dah-Wei Hsu/Heather Orosco
Filer Authorized By:	David Dah-Wei Hsu
Attorney Docket Number:	080015-028410US-1199353
Receipt Date:	27-OCT-2020
Filing Date:	17-AUG-2020
Time Stamp:	16:44:07
Application Type:	Utility under 35 USC 111(a)

Payment information:

Submitted with Payment	no

File Listing:

Document Number	Document Description	File Name	File Size(Bytes)/ Message Digest	Multi Part /.zip	Pages (if appl.)
1		1199353_2nd_Preliminary_Amendment.pdf	95822 30268ae291d5f44c341dbbe69761fb31c3f76b32	yes	5

	Multipart Description/PDF files in .zip description		
	Document Description	**Start**	**End**
	Applicant Arguments/Remarks Made in an Amendment	4	5
	Specification	2	3
	Preliminary Amendment	1	1

Warnings:					
Information:					
2	Sequence Listing (Text File)	080015-028410US-1199353_SL.txt	1709 	no	-
Warnings:					
Information:					
3	Miscellaneous Incoming Letter	1199353_Notice_to_Comply.pdf	150555 c1b7c94e354fe07321f3ae8f5db89bf89bf5c228	no	3
Warnings:					
Information:					
		Total Files Size (in bytes):	248086		

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.

<u>New Applications Under 35 U.S.C. 111</u>
If a new application is being filed and the application includes the necessary components for a 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.
<u>National Stage of an International Application under 35 U.S.C. 371</u>
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.
<u>New International Application Filed with the USPTO as a Receiving Office</u>
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.

REMARKS/ARGUMENTS

The specification has been amended herewith to insert sequence identifiers, to request insertion of sequence listing into the application, and to correct clerical errors. The amendments to the specification contain no new matter.

Applicant submits herewith a sequence listing in computer readable form containing SEQ ID NOS:1-5. The Sequence Listing in computer readable form was prepared in accordance with 37 C.F.R. §§ 1.821 to 1.825. The Sequence Listing does not include new matter or matter that goes beyond the disclosure of the application as filed. According to the Legal Framework for EFS-Web (September 2008), if a sequence listing text file submitted via EFS-Web complies with the requirements of 37 C.F.R. §§ 1.824(a)(2)-(6) and (b), the text file will serve as both the paper copy required by 37 C.F.R. § 1.821(c) and the CRF required by 37 C.F.R. § 1.821(e). Therefore, a paper copy of the referenced Sequence Listing is not included with this amendment.

CONCLUSION

In view of the foregoing, Applicant believes all claims now pending in this application are in condition for examination.

Except for the issue fees payable under 37 C.F.R. § 1.18, the Director is authorized to charge any additional fees during pendency of this application, including any required extension of time fees, or credit any overpayment to Deposit Account Number 20-1430. This paragraph is intended to be a constructive petition for extension of time in accordance with 37 C.F.R. § 1.136(a)(3).

If the Examiner believes a telephone conference would expedite prosecution of this application, please contact the undersigned at (303) 571-4000 or dhsu@kilpatricktownsend.com.

Respectfully submitted,

/David D. Hsu/
David D. Hsu
Registration No. 71659

KILPATRICK TOWNSEND & STOCKTON LLP

DDH:hao

AMENDMENTS TO THE SPECIFICATION

Please add the following new heading and paragraph after paragraph **[0001]**:

REFERENCE TO A "SEQUENCE LISTING" SUBMITTED AS ASCII TEXT FILES VIA EFS-WEB

The instant application contains a Sequence Listing which has been filed electronically in ASCII format and is hereby incorporated by reference in its entirety. Said ASCII copy, created on October 22, 2020, is named 080015-028410US-1199353_SL.txt and is 1,709 bytes in size.

Please replace paragraph **[0044]** with the following amended paragraph:

[0044] **FIG. 31A** shows an example of one approach for generating molecules with unmethylated adenines by whole genome amplification according to embodiments of the present invention. FIG. discloses SEQ ID NO: 4.

Please replace paragraph **[0045]** with the following amended paragraph:

[0045] **FIG. 31B** shows an example of one approach for generating molecules with methylated adenines by whole genome amplification according to embodiments of the present invention. FIG. discloses SEQ ID NO: 5.

Please replace paragraph **[0436]** with the following amended paragraph:

[0436] A first CRISPR/Cas9 complex for introducing a first cut: (all sequences from 5' to 3')

crRNA: GCCUGUAAUCCCAGCACUUUGUUUUAGAGCUAUGCU (SEQ ID NO: 1)

tracrRNA:
AGCAUAGCAAGUUAAAAUAAGGCUAGUCCGUUAUCAACUUGAAAAAGUGGCACC
GAGUCGGUGCUUU (SEQ ID NO: 2)

Please replace paragraph **[0437]** with the following amended paragraph:

[0437] A second CRISPR/Cas9 complex for introducing a second cut:

crRNA: AGGGUCUCGCUCUGUCGCCCGUUUUAGAGCUAUGCU (SEQ ID NO: 3)

tracrRNA:
AGCAUAGCAAGUUAAAAUAAGGCUAGUCCGUUAUCAACUUGAAAAAGUGGCACC
GAGUCGGUGCUUU (SEQ ID NO: 2)

Please insert the SEQUENCE LISTING containing SEQ ID NOS:1-5 at the end of the application.

PATENT APPLICATION FEE DETERMINATION RECORD Substitute for Form PTO-875	Application or Docket Number 16/995,607

APPLICATION AS FILED - PART I

	(Column 1)	(Column 2)	SMALL ENTITY		OR	OTHER THAN SMALL ENTITY	
FOR	NUMBER FILED	NUMBER EXTRA	RATE($)	FEE($)		RATE($)	FEE($)
BASIC FEE (37 CFR 1.16(a), (b), or (c))	N/A	N/A	N/A	75		N/A	
SEARCH FEE (37 CFR 1.16(k), (i), or (m))	N/A	N/A	N/A	330		N/A	
EXAMINATION FEE (37 CFR 1.16(o), (p), or (q))	N/A	N/A	N/A	380		N/A	
TOTAL CLAIMS (37 CFR 1.16(i))	19 minus 20 =	*	x 50 =	0.00	OR		
INDEPENDENT CLAIMS (37 CFR 1.16(h))	1 minus 3 =	*	x 230 =	0.00			
APPLICATION SIZE FEE (37 CFR 1.16(s))	If the specification and drawings exceed 100 sheets of paper, the application size fee due is $310 ($155 for small entity) for each additional 50 sheets or fraction thereof. See 35 U.S.C. 41(a)(1)(G) and 37 CFR 1.16(s).			800			
MULTIPLE DEPENDENT CLAIM PRESENT (37 CFR 1.16(j))				0.00			
* If the difference in column 1 is less than zero, enter "0" in column 2.			TOTAL	1585		TOTAL	

APPLICATION AS AMENDED - PART II

		(Column 1)		(Column 2)	(Column 3)	SMALL ENTITY		OR	OTHER THAN SMALL ENTITY	
		CLAIMS REMAINING AFTER AMENDMENT		HIGHEST NUMBER PREVIOUSLY PAID FOR	PRESENT EXTRA	RATE($)	ADDITIONAL FEE($)		RATE($)	ADDITIONAL FEE($)
AMENDMENT A	Total (37 CFR 1.16(i))	*	Minus	**	=	x =		OR	x =	
	Independent (37 CFR 1.16(h))	*	Minus	***	=	x =		OR	x =	
	Application Size Fee (37 CFR 1.16(s))									
	FIRST PRESENTATION OF MULTIPLE DEPENDENT CLAIM (37 CFR 1.16(j))							OR		
						TOTAL ADD'L FEE		OR	TOTAL ADD'L FEE	

		(Column 1)		(Column 2)	(Column 3)					
		CLAIMS REMAINING AFTER AMENDMENT		HIGHEST NUMBER PREVIOUSLY PAID FOR	PRESENT EXTRA	RATE($)	ADDITIONAL FEE($)		RATE($)	ADDITIONAL FEE($)
AMENDMENT B	Total (37 CFR 1.16(i))	*	Minus	**	=	x =		OR	x =	
	Independent (37 CFR 1.16(h))	*	Minus	***	=	x =		OR	x =	
	Application Size Fee (37 CFR 1.16(s))									
	FIRST PRESENTATION OF MULTIPLE DEPENDENT CLAIM (37 CFR 1.16(j))							OR		
						TOTAL ADD'L FEE		OR	TOTAL ADD'L FEE	

* If the entry in column 1 is less than the entry in column 2, write "0" in column 3.
** If the "Highest Number Previously Paid For" IN THIS SPACE is less than 20, enter "20".
*** If the "Highest Number Previously Paid For" IN THIS SPACE is less than 3, enter "3".
The "Highest Number Previously Paid For" (Total or Independent) is the highest found in the appropriate box in column 1.

UNITED STATES PATENT AND TRADEMARK OFFICE

UNITED STATES DEPARTMENT OF COMMERCE
United States Patent and Trademark Office
Address: COMMISSIONER FOR PATENTS
P.O. Box 1450
Alexandria, Virginia 22313-1450
www.uspto.gov

APPLICATION NUMBER	FILING OR 371(C) DATE	FIRST NAMED APPLICANT	ATTY. DOCKET NO./TITLE
16/995,607	08/17/2020	Yuk-Ming Dennis Lo	080015-028410US-1199353

CONFIRMATION NO. 1256

20350
KILPATRICK TOWNSEND & STOCKTON LLP
Mailstop: IP Docketing - 22
1100 Peachtree Street
Suite 2800
Atlanta, GA 30309

FORMALITIES LETTER

OC000000120167457

Date Mailed: 09/25/2020

NOTICE TO COMPLY WITH REQUIREMENTS FOR PATENT APPLICATIONS CONTAINING NUCLEOTIDE AND/OR AMINO ACID SEQUENCE DISCLOSURES

Filing Date Granted

Applicant is given **TWO MONTHS FROM THE DATE OF THIS NOTICE** within which to comply with the sequence rules, 37 CFR §§ 1.821-1.825. Failure to comply with these requirements will result in ABANDONMENT of the application under 37 CFR §1.821(g). Extension of time may be obtained by filing a petition accompanied by the extension fee under the provisions of 37 CFR §1.136.

Items 1) and 2) provide general guidance related to requirements for sequence disclosures.

1) 37 CFR 1.821(c) requires that patent applications which contain disclosures of nucleotide and/or amino acid sequences that fall within the definitions of 37 CFR 1.821(a) must contain a "Sequence Listing", *as a separate part of the disclosure*, which presents the nucleotide and/or amino acid sequences and associated information using the symbols and format in accordance with the requirements of 37 CFR 1.821-1.825. This "Sequence Listing" part of the disclosure may be submitted:
   a) via EFS-Web (see Section I.1 of the Legal Framework for EFS-Web (https://www.uspto.gov/patents-application-process/filing-online/legal-framework-efs-web), hereinafter "Legal Framework") as an ASCII text file, together with an incorporation-by-reference of the material in the ASCII text file in a separate paragraph of the specification identifying:
      i) the name of the ASCII text file;
      ii) the date of creation; and
      iii) the size of the ASCII text file in bytes;
   b) on compact disc(s) in duplicate according to 37 CFR 1.52(e)(1)(ii) and (4), labeled according to 37 CFR 1.52(e)(6), together with a statement that the duplicate compact discs are identical and an incorporation-by-reference of the material in the ASCII text file according to 37 CFR 1.52(e)(5) in a separate paragraph of the specification identifying:
      i) the name of the ASCII text file;
      ii) the date of creation; and
      iii) the size of the ASCII text file in bytes;
   c) via EFS-Web as a PDF (not recommended); or
   d) on paper.
2) 37 CFR 1.821(e) requires that a *copy* of the "Sequence Listing" must also be submitted in computer readable form (CRF) in accordance with the requirements of 37 CFR 1.824.
   a) If a "Sequence Listing" ASCII text file submitted via EFS-Web on the application filing date complies with the requirements of 37 CFR 1.824(a)(2)-(6) and (b), and applicant has not filed a "Sequence Listing" as

a PDF, the text file will serve as both the "Sequence Listing" required by 37 CFR 1.821(c) and the CRF required by 37 CFR 1.821(e), and the statement of identity under the "Legal Framework" is not required.

b) If the "Sequence Listing" required by 37 CFR 1.821(c) is filed via EFS-Web as a PDF, then the "Legal Framework" requires submission of a statement that the "Sequence Listing" content of the PDF copy and the ASCII text file copy submitted via EFS-Web are identical.

c) If the "Sequence Listing" required by 37 CFR 1.821(c) is filed on paper or compact disc, then 37 CFR 1.821(f) requires submission of a statement that the "Sequence Listing" content of the paper or compact disc copy and the CRF are identical.

Specific deficiencies and the required response to this notice are as follows:

- This application fails to comply with the requirements of 37 CFR 1.821-1.825 because it does not contain a "Sequence Listing" as a separate part of the disclosure or a CRF. Applicant must provide:
  - A "Sequence Listing" part of the disclosure, as described above in item 1); **together with**
  - **An amendment specifically directing its entry into the application;**
  - A statement that the "Sequence Listing" includes no new matter; and
  - A statement that indicates support for the amendment in the application, as filed, as required by 37 CFR 1.825 and the "Legal Framework".
- If the "Sequence Listing" part of the disclosure is submitted according to item 1) a) or b) above, Applicant must also provide:
  - A substitute specification in compliance with 37 CFR 1.121(b)(3) and 1.125 inserting the required incorporation-by-reference paragraph, consisting of:
    - A copy of the previously-submitted specification, with deletions shown with strikethrough or brackets and insertions shown with underlining (marked-up version);
    - A copy of the amended specification without markings (clean version); and
    - A statement that the substitute specification contains no new matter.
- If the "Sequence Listing" part of the disclosure is submitted according to item 1) b), c), or d) above, Applicant must also provide:
  - A CRF; and
  - A statement according to item 2) b) or c) above.
- If applicant desires the CRF in this application to be identical with the CRF of another application on file in the U.S. Patent and Trademark Office, such request in accordance with 37 CFR 1.821(e) may be submitted in lieu of a new CRF (Form PTO/SB/93 may be used for this purpose).

For questions regarding compliance to these requirements, please contact:

- **For Rules interpretation, call (571) 272-2510**
- **For CRF submission help, call (571) 272-2510**
- **For PatentIn software program support technical assistance, call (866) 217-9197.**
- **Send e-mail correspondence for PatentIn software program help to ebc@uspto.gov**
- **PatentIn software is available at http://www.uspto.gov/patents/resources/tools/checker/patentinrel.jsp .**

Applicant is cautioned that correction of the above items may cause the specification and drawings page count to exceed 100 pages. If the specification and drawings exceed 100 pages, applicant will need to submit the required application size fee.

Replies must be received in the USPTO within the set time period or must include a proper Certificate of Mailing or Transmission under 37 CFR 1.8 with a mailing or transmission date within the set time period. For more information and a suggested format, see Form PTO/SB/92 and MPEP 512.

Replies should be mailed to:

Mail Stop Missing Parts
Commissioner for Patents
P.O. Box 1450
Alexandria VA 22313-1450

Registered users of EFS-Web may alternatively submit their reply to this notice via EFS-Web, including a copy of this Notice and selecting the document description "Applicant response to Pre-Exam Formalities Notice". https://sportal.uspto.gov/authenticate/AuthenticateUserLocalEPF.html

For more information about EFS-Web please call the USPTO Electronic Business Center at 1-866-217-9197 or visit our website at http://www.uspto.gov/ebc.

If you are not using EFS-Web to submit your reply, you must include a copy of this notice.

Questions about the contents of this notice and the requirements it sets forth should be directed to the Office of Data Management, Application Assistance Unit, at **(571) 272-4000** or **(571) 272-4200** or **1-888-786-0101**.

/sareebuddin/

UNITED STATES PATENT AND TRADEMARK OFFICE

UNITED STATES DEPARTMENT OF COMMERCE
United States Patent and Trademark Office
Address: COMMISSIONER FOR PATENTS
P.O. Box 1450
Alexandria, Virginia 22313-1450
www.uspto.gov

APPLICATION NUMBER	FILING or 371(c) DATE	GRP ART UNIT	FIL FEE REC'D	ATTY.DOCKET.NO	TOT CLAIMS	IND CLAIMS
16/995,607	08/17/2020		1585	080015-028410US-1199353	19	1

CONFIRMATION NO. 1256

FILING RECEIPT

20350
KILPATRICK TOWNSEND & STOCKTON LLP
Mailstop: IP Docketing - 22
1100 Peachtree Street
Suite 2800
Atlanta, GA 30309

OC000000120167456

Date Mailed: 09/25/2020

Receipt is acknowledged of this non-provisional utility patent application. The application will be taken up for examination in due course. Applicant will be notified as to the results of the examination. Any correspondence concerning the application must include the following identification information: the U.S. APPLICATION NUMBER, FILING DATE, NAME OF FIRST INVENTOR, and TITLE OF INVENTION. Fees transmitted by check or draft are subject to collection.

Please verify the accuracy of the data presented on this receipt. If an error is noted on this Filing Receipt, please submit a written request for a corrected Filing Receipt, including a properly marked-up ADS showing the changes with strike-through for deletions and underlining for additions. If you received a "Notice to File Missing Parts" or other Notice requiring a response for this application, please submit any request for correction to this Filing Receipt with your reply to the Notice. When the USPTO processes the reply to the Notice, the USPTO will generate another Filing Receipt incorporating the requested corrections provided that the request is grantable.

Inventor(s)

Yuk-Ming Dennis Lo, Homantin, Kowloon, Hong Kong SAR, CHINA;
Rossa Wai Kwun Chiu, Shatin, New Territories, Hong Kong SAR, CHINA;
Kwan Chee Chan, Jordan, Hong Kong SAR, CHINA;
Peiyong Jiang, Shatin, New Territories, Hong Kong SAR, CHINA;
Suk Hang Cheng, Fanling, New Territories, Hong Kong SAR, CHINA;
Wenlei Peng, Shatin, New Territories, Hong Kong SAR, CHINA;
On Yee Tse, Fanling, New Territories, Hong Kong SAR, CHINA;

Applicant(s)

The Chinese University of Hong Kong, Shatin, HONG KONG;

Power of Attorney: The patent practitioners associated with Customer Number 20350

Domestic Priority data as claimed by applicant

This appln claims benefit of 63/051,210 07/13/2020
and claims benefit of 63/019,790 05/04/2020
and claims benefit of 62/991,891 03/19/2020
and claims benefit of 62/970,586 02/05/2020
and claims benefit of 62/887,987 08/16/2019

Foreign Applications for which priority is claimed (You may be eligible to benefit from the **Patent Prosecution Highway** program at the USPTO. Please see http://www.uspto.gov for more information.) - None.
Foreign application information must be provided in an Application Data Sheet in order to constitute a claim to foreign priority. See 37 CFR 1.55 and 1.76.

Permission to Access Application via Priority Document Exchange: Yes

Permission to Access Search Results: Yes

Applicant may provide or rescind an authorization for access using Form PTO/SB/39 or Form PTO/SB/69 as appropriate.

If Required, Foreign Filing License Granted: 09/24/2020

The country code and number of your priority application, to be used for filing abroad under the Paris Convention, is **US 16/995,607**

Projected Publication Date: To Be Determined - pending completion of Sequence Disclosures

Non-Publication Request: No

Early Publication Request: No

**** SMALL ENTITY ****

Title

DETERMINATION OF BASE MODIFICATIONS OF NUCLEIC ACIDS

Preliminary Class

Statement under 37 CFR 1.55 or 1.78 for AIA (First Inventor to File) Transition Applications: No

PROTECTING YOUR INVENTION OUTSIDE THE UNITED STATES

Since the rights granted by a U.S. patent extend only throughout the territory of the United States and have no effect in a foreign country, an inventor who wishes patent protection in another country must apply for a patent in a specific country or in regional patent offices. Applicants may wish to consider the filing of an international application under the Patent Cooperation Treaty (PCT). An international (PCT) application generally has the same effect as a regular national patent application in each PCT-member country. The PCT process **simplifies** the filing of patent applications on the same invention in member countries, but **does not result** in a grant of "an international patent" and does not eliminate the need of applicants to file additional documents and fees in countries where patent protection is desired.

Almost every country has its own patent law, and a person desiring a patent in a particular country must make an application for patent in that country in accordance with its particular laws. Since the laws of many countries differ in various respects from the patent law of the United States, applicants are advised to seek guidance from specific foreign countries to ensure that patent rights are not lost prematurely.

Applicants also are advised that in the case of inventions made in the United States, the Director of the USPTO must issue a license before applicants can apply for a patent in a foreign country. The filing of a U.S. patent application serves as a request for a foreign filing license. The application's filing receipt contains further information and guidance as to the status of applicant's license for foreign filing.

Applicants may wish to consult the USPTO booklet, "General Information Concerning Patents" (specifically, the section entitled "Treaties and Foreign Patents") for more information on timeframes and deadlines for filing foreign patent applications. The guide is available either by contacting the USPTO Contact Center at 800-786-9199, or it can be viewed on the USPTO website at http://www.uspto.gov/web/offices/pac/doc/general/index.html.

For information on preventing theft of your intellectual property (patents, trademarks and copyrights), you may wish to consult the U.S. Government website, http://www.stopfakes.gov. Part of a Department of Commerce initiative, this website includes self-help "toolkits" giving innovators guidance on how to protect intellectual property in specific countries such as China, Korea and Mexico. For questions regarding patent enforcement issues, applicants may call the U.S. Government hotline at 1-866-999-HALT (1-866-999-4258).

LICENSE FOR FOREIGN FILING UNDER

Title 35, United States Code, Section 184

Title 37, Code of Federal Regulations, 5.11 & 5.15

GRANTED

The applicant has been granted a license under 35 U.S.C. 184, if the phrase "IF REQUIRED, FOREIGN FILING LICENSE GRANTED" followed by a date appears on this form. Such licenses are issued in all applications where the conditions for issuance of a license have been met, regardless of whether or not a license may be required as set forth in 37 CFR 5.15. The scope and limitations of this license are set forth in 37 CFR 5.15(a) unless an earlier license has been issued under 37 CFR 5.15(b). The license is subject to revocation upon written notification. The date indicated is the effective date of the license, unless an earlier license of similar scope has been granted under 37 CFR 5.13 or 5.14.

This license is to be retained by the licensee and may be used at any time on or after the effective date thereof unless it is revoked. This license is automatically transferred to any related applications(s) filed under 37 CFR 1.53(d). This license is not retroactive.

The grant of a license does not in any way lessen the responsibility of a licensee for the security of the subject matter as imposed by any Government contract or the provisions of existing laws relating to espionage and the national security or the export of technical data. Licensees should apprise themselves of current regulations especially with respect to certain countries, of other agencies, particularly the Office of Defense Trade Controls, Department of State (with respect to Arms, Munitions and Implements of War (22 CFR 121-128)); the Bureau of Industry and Security, Department of Commerce (15 CFR parts 730-774); the Office of Foreign AssetsControl, Department of Treasury (31 CFR Parts 500+) and the Department of Energy.

NOT GRANTED

No license under 35 U.S.C. 184 has been granted at this time, if the phrase "IF REQUIRED, FOREIGN FILING LICENSE GRANTED" DOES NOT appear on this form. Applicant may still petition for a license under 37 CFR 5.12, if a license is desired before the expiration of 6 months from the filing date of the application. If 6 months has lapsed from the filing date of this application and the licensee has not received any indication of a secrecy order under 35 U.S.C. 181, the licensee may foreign file the application pursuant to 37 CFR 5.15(b).

SelectUSA

The United States represents the largest, most dynamic marketplace in the world and is an unparalleled location for business investment, innovation, and commercialization of new technologies. The U.S. offers tremendous resources and advantages for those who invest and manufacture goods here. Through SelectUSA, our nation works to promote and facilitate business investment. SelectUSA provides information assistance to the international investor

community; serves as an ombudsman for existing and potential investors; advocates on behalf of U.S. cities, states, and regions competing for global investment; and counsels U.S. economic development organizations on investment attraction best practices. To learn more about why the United States is the best country in the world to develop technology, manufacture products, deliver services, and grow your business, visit http://www.SelectUSA.gov or call +1-202-482-6800.

United States Patent and Trademark Office

Office of the Chief Financial Officer

Document Code:WFEE

User :C43106

Sale Accounting Date:09/24/2020

Sale Item Reference Number	Effective Date
16995607	08/17/2020

Document Number	Fee Code	Fee Code Description	Amount Paid	Payment Method
I20209NA27431466	2081	UTILITY APPL SIZE FEE PER 50 SHEETS >100	$200.00	Deposit Account

PTO/AIA/01 (06-12)
Approved for use through 01/31/2014. OMB 0651-0032
U.S. Patent and Trademark Office; U.S. DEPARTMENT OF COMMERCE
Under the Paperwork Reduction Act of 1995, no persons are required to respond to a collection of information unless it displays a valid OMB control number.

DECLARATION (37 CFR 1.63) FOR UTILITY OR DESIGN APPLICATION USING AN APPLICATION DATA SHEET (37 CFR 1.76)

Title of Invention	DETERMINATION OF BASE MODIFICATIONS OF NUCLEIC ACIDS

As the below named inventor, I hereby declare that:

This declaration is directed to: ☐ The attached application, or

☒ United States application or PCT international application number 16/995,607

filed on August 17, 2020

The above-identified application was made or authorized to be made by me.

I believe that I am the original inventor or an original joint inventor of a claimed invention in the application.

I hereby acknowledge that any willful false statement made in this declaration is punishable under 18 U.S.C. 1001 by fine or imprisonment of not more than five (5) years, or both.

WARNING:

Petitioner/applicant is cautioned to avoid submitting personal information in documents filed in a patent application that may contribute to identity theft. Personal information such as social security numbers, bank account numbers, or credit card numbers (other than a check or credit card authorization form PTO-2038 submitted for payment purposes) is never required by the USPTO to support a petition or an application. If this type of personal information is included in documents submitted to the USPTO, petitioners/applicants should consider redacting such personal information from the documents before submitting them to the USPTO. Petitioner/applicant is advised that the record of a patent application is available to the public after publication of the application (unless a non-publication request in compliance with 37 CFR 1.213(a) is made in the application) or issuance of a patent. Furthermore, the record from an abandoned application may also be available to the public if the application is referenced in a published application or an issued patent (see 37 CFR 1.14). Checks and credit card authorization forms PTO-2038 submitted for payment purposes are not retained in the application file and therefore are not publicly available.

LEGAL NAME OF INVENTOR

Inventor: Yuk-Ming Dennis Lo Date (Optional) : _____

Signature: _____

Note: An application data sheet (PTO/SB/14 or equivalent), including naming the entire inventive entity, must accompany this form. Use an additional PTO/AIA/01 form for each additional inventor.

This collection of information is required by 35 U.S.C. 115 and 37 CFR 1.63. The information is required to obtain or retain a benefit by the public which is to file (and by the USPTO to process) an application. Confidentiality is governed by 35 U.S.C. 122 and 37 CFR 1.11 and 1.14. This collection is estimated to take 1 minute to complete, including gathering, preparing, and submitting the completed application form to the USPTO. Time will vary depending upon the individual case. Any comments on the amount of time you require to complete this form and/or suggestions for reducing this burden, should be sent to the Chief Information Officer, U.S. Patent and Trademark Office, U.S. Department of Commerce, P.O. Box 1450, Alexandria, VA 22313-1450. DO NOT SEND FEES OR COMPLETED FORMS TO THIS ADDRESS. SEND TO: Commissioner for Patents, P.O. Box 1450, Alexandria, VA 22313-1450.

If you need assistance in completing the form, call 1-800-PTO-9199 and select option 2.

PTO/AIA/01 (06-12)
Approved for use through 01/31/2014. OMB 0651-0032
U.S. Patent and Trademark Office; U.S. DEPARTMENT OF COMMERCE
Under the Paperwork Reduction Act of 1995, no persons are required to respond to a collection of information unless it displays a valid OMB control number.

DECLARATION (37 CFR 1.63) FOR UTILITY OR DESIGN APPLICATION USING AN APPLICATION DATA SHEET (37 CFR 1.76)

Title of Invention	DETERMINATION OF BASE MODIFICATIONS OF NUCLEIC ACIDS

As the below named inventor, I hereby declare that:

This declaration is directed to: ☐ The attached application, or

☒ United States application or PCT international application number 16/995,607

filed on August 17, 2020

The above-identified application was made or authorized to be made by me.

I believe that I am the original inventor or an original joint inventor of a claimed invention in the application.

I hereby acknowledge that any willful false statement made in this declaration is punishable under 18 U.S.C. 1001 by fine or imprisonment of not more than five (5) years, or both.

WARNING:

Petitioner/applicant is cautioned to avoid submitting personal information in documents filed in a patent application that may contribute to identity theft. Personal information such as social security numbers, bank account numbers, or credit card numbers (other than a check or credit card authorization form PTO-2038 submitted for payment purposes) is never required by the USPTO to support a petition or an application. If this type of personal information is included in documents submitted to the USPTO, petitioners/applicants should consider redacting such personal information from the documents before submitting them to the USPTO. Petitioner/applicant is advised that the record of a patent application is available to the public after publication of the application (unless a non-publication request in compliance with 37 CFR 1.213(a) is made in the application) or issuance of a patent. Furthermore, the record from an abandoned application may also be available to the public if the application is referenced in a published application or an issued patent (see 37 CFR 1.14). Checks and credit card authorization forms PTO-2038 submitted for payment purposes are not retained in the application file and therefore are not publicly available.

LEGAL NAME OF INVENTOR

Inventor: Rossa Wai Kwun Chiu Date (Optional) :

Signature:

Note: An application data sheet (PTO/SB/14 or equivalent), including naming the entire inventive entity, must accompany this form. Use an additional PTO/AIA/01 form for each additional inventor.

This collection of information is required by 35 U.S.C. 115 and 37 CFR 1.63. The information is required to obtain or retain a benefit by the public which is to file (and by the USPTO to process) an application. Confidentiality is governed by 35 U.S.C. 122 and 37 CFR 1.11 and 1.14. This collection is estimated to take 1 minute to complete, including gathering, preparing, and submitting the completed application form to the USPTO. Time will vary depending upon the individual case. Any comments on the amount of time you require to complete this form and/or suggestions for reducing this burden, should be sent to the Chief Information Officer, U.S. Patent and Trademark Office, U.S. Department of Commerce, P.O. Box 1450, Alexandria, VA 22313-1450. DO NOT SEND FEES OR COMPLETED FORMS TO THIS ADDRESS. SEND TO: Commissioner for Patents, P.O. Box 1450, Alexandria, VA 22313-1450.

If you need assistance in completing the form, call 1-800-PTO-9199 and select option 2.

PTO/AIA/01 (06-12)
Approved for use through 01/31/2014. OMB 0651-0032
U.S. Patent and Trademark Office; U.S. DEPARTMENT OF COMMERCE
Under the Paperwork Reduction Act of 1995, no persons are required to respond to a collection of information unless it displays a valid OMB control number.

DECLARATION (37 CFR 1.63) FOR UTILITY OR DESIGN APPLICATION USING AN APPLICATION DATA SHEET (37 CFR 1.76)

Title of Invention	DETERMINATION OF BASE MODIFICATIONS OF NUCLEIC ACIDS

As the below named inventor, I hereby declare that:

This declaration is directed to: ☐ The attached application, or

☑ United States application or PCT international application number 16/995,607

filed on August 17, 2020

The above-identified application was made or authorized to be made by me.

I believe that I am the original inventor or an original joint inventor of a claimed invention in the application.

I hereby acknowledge that any willful false statement made in this declaration is punishable under 18 U.S.C. 1001 by fine or imprisonment of not more than five (5) years, or both.

WARNING:

Petitioner/applicant is cautioned to avoid submitting personal information in documents filed in a patent application that may contribute to identity theft. Personal information such as social security numbers, bank account numbers, or credit card numbers (other than a check or credit card authorization form PTO-2038 submitted for payment purposes) is never required by the USPTO to support a petition or an application. If this type of personal information is included in documents submitted to the USPTO, petitioners/applicants should consider redacting such personal information from the documents before submitting them to the USPTO. Petitioner/applicant is advised that the record of a patent application is available to the public after publication of the application (unless a non-publication request in compliance with 37 CFR 1.213(a) is made in the application) or issuance of a patent. Furthermore, the record from an abandoned application may also be available to the public if the application is referenced in a published application or an issued patent (see 37 CFR 1.14). Checks and credit card authorization forms PTO-2038 submitted for payment purposes are not retained in the application file and therefore are not publicly available.

LEGAL NAME OF INVENTOR

Inventor: Kwan Chee Chan Date (Optional): _____

Signature: _____

Note. An application data sheet (PTO/SB/14 or equivalent), including naming the entire inventive entity, must accompany this form. Use an additional PTO/AIA/01 form for each additional inventor.

This collection of information is required by 35 U.S.C. 115 and 37 CFR 1.63. The information is required to obtain or retain a benefit by the public which is to file (and by the USPTO to process) an application. Confidentiality is governed by 35 U.S.C. 122 and 37 CFR 1.11 and 1.14. This collection is estimated to take 1 minute to complete, including gathering, preparing, and submitting the completed application form to the USPTO. Time will vary depending upon the individual case. Any comments on the amount of time you require to complete this form and/or suggestions for reducing this burden, should be sent to the Chief Information Officer, U.S. Patent and Trademark Office, U.S. Department of Commerce, P.O. Box 1450, Alexandria, VA 22313-1450. DO NOT SEND FEES OR COMPLETED FORMS TO THIS ADDRESS. SEND TO: Commissioner for Patents, P.O. Box 1450, Alexandria, VA 22313-1450.

If you need assistance in completing the form, call 1-800-PTO-9199 and select option 2.

PTO/AIA/01 (06-12)
Approved for use through 01/31/2014. OMB 0651-0032
U.S. Patent and Trademark Office; U.S. DEPARTMENT OF COMMERCE
Under the Paperwork Reduction Act of 1995, no persons are required to respond to a collection of information unless it displays a valid OMB control number.

DECLARATION (37 CFR 1.63) FOR UTILITY OR DESIGN APPLICATION USING AN APPLICATION DATA SHEET (37 CFR 1.76)

Title of Invention	DETERMINATION OF BASE MODIFICATIONS OF NUCLEIC ACIDS

As the below named inventor, I hereby declare that:

This declaration is directed to: ☐ The attached application, or

☒ United States application or PCT international application number 16/995,607

filed on August 17, 2020

The above-identified application was made or authorized to be made by me.

I believe that I am the original inventor or an original joint inventor of a claimed invention in the application.

I hereby acknowledge that any willful false statement made in this declaration is punishable under 18 U.S.C. 1001 by fine or imprisonment of not more than five (5) years, or both.

WARNING:

Petitioner/applicant is cautioned to avoid submitting personal information in documents filed in a patent application that may contribute to identity theft. Personal information such as social security numbers, bank account numbers, or credit card numbers (other than a check or credit card authorization form PTO-2038 submitted for payment purposes) is never required by the USPTO to support a petition or an application. If this type of personal information is included in documents submitted to the USPTO, petitioners/applicants should consider redacting such personal information from the documents before submitting them to the USPTO. Petitioner/applicant is advised that the record of a patent application is available to the public after publication of the application (unless a non-publication request in compliance with 37 CFR 1.213(a) is made in the application) or issuance of a patent. Furthermore, the record from an abandoned application may also be available to the public if the application is referenced in a published application or an issued patent (see 37 CFR 1.14). Checks and credit card authorization forms PTO-2038 submitted for payment purposes are not retained in the application file and therefore are not publicly available.

LEGAL NAME OF INVENTOR

Inventor: Peiyong Jiang Date (Optional): _____

Signature: Jiang Peiyong

Note: An application data sheet (PTO/SB/14 or equivalent), including naming the entire inventive entity, must accompany this form. Use an additional PTO/AIA/01 form for each additional inventor.

This collection of information is required by 35 U.S.C. 115 and 37 CFR 1.63. The information is required to obtain or retain a benefit by the public which is to file (and by the USPTO to process) an application. Confidentiality is governed by 35 U.S.C. 122 and 37 CFR 1.11 and 1.14. This collection is estimated to take 1 minute to complete, including gathering, preparing, and submitting the completed application form to the USPTO. Time will vary depending upon the individual case. Any comments on the amount of time you require to complete this form and/or suggestions for reducing this burden, should be sent to the Chief Information Officer, U.S. Patent and Trademark Office, U.S. Department of Commerce, P.O. Box 1450, Alexandria, VA 22313-1450. DO NOT SEND FEES OR COMPLETED FORMS TO THIS ADDRESS. SEND TO: Commissioner for Patents, P.O. Box 1450, Alexandria, VA 22313-1450.

If you need assistance in completing the form, call 1-800-PTO-9199 and select option 2.

PTO/AIA/01 (06-12)
Approved for use through 01/31/2014. OMB 0651-0032
U.S. Patent and Trademark Office; U.S. DEPARTMENT OF COMMERCE
Under the Paperwork Reduction Act of 1995, no persons are required to respond to a collection of information unless it displays a valid OMB control number.

DECLARATION (37 CFR 1.63) FOR UTILITY OR DESIGN APPLICATION USING AN APPLICATION DATA SHEET (37 CFR 1.76)

Title of Invention	DETERMINATION OF BASE MODIFICATIONS OF NUCLEIC ACIDS

As the below named inventor, I hereby declare that:

This declaration is directed to: ☐ The attached application, or

☒ United States application or PCT international application number 16/995,607

filed on August 17, 2020

The above-identified application was made or authorized to be made by me.

I believe that I am the original inventor or an original joint inventor of a claimed invention in the application.

I hereby acknowledge that any willful false statement made in this declaration is punishable under 18 U.S.C. 1001 by fine or imprisonment of not more than five (5) years, or both.

WARNING:

Petitioner/applicant is cautioned to avoid submitting personal information in documents filed in a patent application that may contribute to identity theft. Personal information such as social security numbers, bank account numbers, or credit card numbers (other than a check or credit card authorization form PTO-2038 submitted for payment purposes) is never required by the USPTO to support a petition or an application. If this type of personal information is included in documents submitted to the USPTO, petitioners/applicants should consider redacting such personal information from the documents before submitting them to the USPTO. Petitioner/applicant is advised that the record of a patent application is available to the public after publication of the application (unless a non-publication request in compliance with 37 CFR 1.213(a) is made in the application) or issuance of a patent. Furthermore, the record from an abandoned application may also be available to the public if the application is referenced in a published application or an issued patent (see 37 CFR 1.14). Checks and credit card authorization forms PTO-2038 submitted for payment purposes are not retained in the application file and therefore are not publicly available.

LEGAL NAME OF INVENTOR

Inventor: Suk Hang Cheng Date (Optional) : _____

Signature: _____

Note: An application data sheet (PTO/SB/14 or equivalent), including naming the entire inventive entity, must accompany this form. Use an additional PTO/AIA/01 form for each additional inventor.

This collection of information is required by 35 U.S.C. 115 and 37 CFR 1.63. The information is required to obtain or retain a benefit by the public which is to file (and by the USPTO to process) an application. Confidentiality is governed by 35 U.S.C. 122 and 37 CFR 1.11 and 1.14. This collection is estimated to take 1 minute to complete, including gathering, preparing, and submitting the completed application form to the USPTO. Time will vary depending upon the individual case. Any comments on the amount of time you require to complete this form and/or suggestions for reducing this burden, should be sent to the Chief Information Officer, U.S. Patent and Trademark Office, U.S. Department of Commerce, P.O. Box 1450, Alexandria, VA 22313-1450. DO NOT SEND FEES OR COMPLETED FORMS TO THIS ADDRESS. SEND TO: Commissioner for Patents, P.O. Box 1450, Alexandria, VA 22313-1450.

If you need assistance in completing the form, call 1-800-PTO-9199 and select option 2.

PTO/AIA/01 (06-12)
Approved for use through 01/31/2014. OMB 0651-0032
U.S. Patent and Trademark Office; U.S. DEPARTMENT OF COMMERCE
Under the Paperwork Reduction Act of 1995, no persons are required to respond to a collection of information unless it displays a valid OMB control number.

DECLARATION (37 CFR 1.63) FOR UTILITY OR DESIGN APPLICATION USING AN APPLICATION DATA SHEET (37 CFR 1.76)

Title of Invention	DETERMINATION OF BASE MODIFICATIONS OF NUCLEIC ACIDS

As the below named inventor, I hereby declare that:

This declaration is directed to: ☐ The attached application, or

☒ United States application or PCT international application number 16/995,607

filed on August 17, 2020

The above-identified application was made or authorized to be made by me.

I believe that I am the original inventor or an original joint inventor of a claimed invention in the application.

I hereby acknowledge that any willful false statement made in this declaration is punishable under 18 U.S.C. 1001 by fine or imprisonment of not more than five (5) years, or both.

WARNING:

Petitioner/applicant is cautioned to avoid submitting personal information in documents filed in a patent application that may contribute to identity theft. Personal information such as social security numbers, bank account numbers, or credit card numbers (other than a check or credit card authorization form PTO-2038 submitted for payment purposes) is never required by the USPTO to support a petition or an application. If this type of personal information is included in documents submitted to the USPTO, petitioners/applicants should consider redacting such personal information from the documents before submitting them to the USPTO. Petitioner/applicant is advised that the record of a patent application is available to the public after publication of the application (unless a non-publication request in compliance with 37 CFR 1.213(a) is made in the application) or issuance of a patent. Furthermore, the record from an abandoned application may also be available to the public if the application is referenced in a published application or an issued patent (see 37 CFR 1.14). Checks and credit card authorization forms PTO-2038 submitted for payment purposes are not retained in the application file and therefore are not publicly available.

LEGAL NAME OF INVENTOR

Inventor: Wenlei Peng Date (Optional): Aug 14 2020

Signature: Wenlei Peng

Note: An application data sheet (PTO/SB/14 or equivalent), including naming the entire inventive entity, must accompany this form. Use an additional PTO/AIA/01 form for each additional inventor.

This collection of information is required by 35 U.S.C. 115 and 37 CFR 1.63. The information is required to obtain or retain a benefit by the public which is to file (and by the USPTO to process) an application. Confidentiality is governed by 35 U.S.C. 122 and 37 CFR 1.11 and 1.14. This collection is estimated to take 1 minute to complete, including gathering, preparing, and submitting the completed application form to the USPTO. Time will vary depending upon the individual case. Any comments on the amount of time you require to complete this form and/or suggestions for reducing this burden, should be sent to the Chief Information Officer, U.S. Patent and Trademark Office, U.S. Department of Commerce, P.O. Box 1450, Alexandria, VA 22313-1450. DO NOT SEND FEES OR COMPLETED FORMS TO THIS ADDRESS. SEND TO: Commissioner for Patents, P.O. Box 1450, Alexandria, VA 22313-1450.

If you need assistance in completing the form, call 1-800-PTO-9199 and select option 2.

PTO/AIA/01 (06-12)
Approved for use through 01/31/2014. OMB 0651-0032
U.S. Patent and Trademark Office, U.S. DEPARTMENT OF COMMERCE
Under the Paperwork Reduction Act of 1995, no persons are required to respond to a collection of information unless it displays a valid OMB control number.

DECLARATION (37 CFR 1.63) FOR UTILITY OR DESIGN APPLICATION USING AN APPLICATION DATA SHEET (37 CFR 1.76)

Title of Invention	DETERMINATION OF BASE MODIFICATIONS OF NUCLEIC ACIDS

As the below named inventor, I hereby declare that:

This declaration is directed to: ☐ The attached application, or

☒ United States application or PCT international application number 16/995,607

filed on August 17, 2020

The above-identified application was made or authorized to be made by me.

I believe that I am the original inventor or an original joint inventor of a claimed invention in the application.

I hereby acknowledge that any willful false statement made in this declaration is punishable under 18 U.S.C. 1001 by fine or imprisonment of not more than five (5) years, or both.

WARNING:

Petitioner/applicant is cautioned to avoid submitting personal information in documents filed in a patent application that may contribute to identity theft. Personal information such as social security numbers, bank account numbers, or credit card numbers (other than a check or credit card authorization form PTO-2038 submitted for payment purposes) is never required by the USPTO to support a petition or an application. If this type of personal information is included in documents submitted to the USPTO, petitioners/applicants should consider redacting such personal information from the documents before submitting them to the USPTO. Petitioner/applicant is advised that the record of a patent application is available to the public after publication of the application (unless a non-publication request in compliance with 37 CFR 1.213(a) is made in the application) or issuance of a patent. Furthermore, the record from an abandoned application may also be available to the public if the application is referenced in a published application or an issued patent (see 37 CFR 1.14). Checks and credit card authorization forms PTO-2038 submitted for payment purposes are not retained in the application file and therefore are not publicly available.

LEGAL NAME OF INVENTOR

Inventor: On Yee Tse Date (Optional) : _____

Signature: On Yee

Note: An application data sheet (PTO/SB/14 or equivalent), including naming the entire inventive entity, must accompany this form. Use an additional PTO/AIA/01 form for each additional inventor.

This collection of information is required by 35 U.S.C. 115 and 37 CFR 1.63. The information is required to obtain or retain a benefit by the public which is to file (and by the USPTO to process) an application. Confidentiality is governed by 35 U.S.C. 122 and 37 CFR 1.11 and 1.14. This collection is estimated to take 1 minute to complete, including gathering, preparing, and submitting the completed application form to the USPTO. Time will vary depending upon the individual case. Any comments on the amount of time you require to complete this form and/or suggestions for reducing this burden, should be sent to the Chief Information Officer, U.S. Patent and Trademark Office, U.S. Department of Commerce, P.O. Box 1450, Alexandria, VA 22313-1450. DO NOT SEND FEES OR COMPLETED FORMS TO THIS ADDRESS. SEND TO: Commissioner for Patents, P.O. Box 1450, Alexandria, VA 22313-1450.

If you need assistance in completing the form, call 1-800-PTO-9199 and select option 2.

Electronic Acknowledgement Receipt

EFS ID:	40576695
Application Number:	16995607
International Application Number:	
Confirmation Number:	1256
Title of Invention:	DETERMINATION OF BASE MODIFICATIONS OF NUCLEIC ACIDS
First Named Inventor/Applicant Name:	Yuk-Ming Dennis Lo
Customer Number:	20350
Filer:	David Dah-Wei Hsu/Heather Orosco
Filer Authorized By:	David Dah-Wei Hsu
Attorney Docket Number:	080015-028410US-1199353
Receipt Date:	16-SEP-2020
Filing Date:	
Time Stamp:	15:29:50
Application Type:	Utility under 35 USC 111(a)

Payment information:

Submitted with Payment	no

File Listing:

Document Number	Document Description	File Name	File Size(Bytes)/ Message Digest	Multi Part /.zip	Pages (if appl.)
1	Oath or Declaration filed	1199353_Declarations.pdf	10715586 7db14cad09ac81613a95213b7f2a5f1ef4e137eb	no	7

Warnings:

Information:	
Total Files Size (in bytes):	10715586

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 a 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.

SCORE Placeholder Sheet for IFW Content

Application Number: 16995607 Document Date: 08/17/2020

The presence of this form in the IFW record indicates that the following document type was received in electronic format on the date identified above. This content is stored in the SCORE database.

Since this was an electronic submission, there is no physical artifact folder, no artifact folder is recorded in PALM, and no paper documents or physical media exist. The TIFF images in the IFW record were created from the original documents that are stored in SCORE.

- Drawing

At the time of document entry (noted above):

- USPTO employees may access SCORE content via DAV or via the SCORE web page.
- External customers may access SCORE content via PAIR using the Supplemental Content tab.

Form Revision Date: March 1, 2019

PTO/AIA/80 (07-12)
Approved for use through 11/30/2014. OMB 0651-0035
U.S. Patent and Trademark Office; U.S. DEPARTMENT OF COMMERCE
Under the Paperwork Reduction Act of 1995, no persons are required to respond to a collection of information unless it displays a valid OMB control number.

POWER OF ATTORNEY TO PROSECUTE APPLICATIONS BEFORE THE USPTO

I hereby revoke all previous powers of attorney given in the application identified in the attached statement under 37 CFR 3.73(c).

I hereby appoint:

☒ Practitioners associated with Customer Number: **20350**

OR

☐ Practitioner(s) named below (if more than ten patent practitioners are to be named, then a customer number must be used):

Name	Registration Number	Name	Registration Number

As attorney(s) or agent(s) to represent the undersigned before the United States Patent and Trademark Office (USPTO) in connection with any and all patent applications assigned only to the undersigned according to the USPTO assignment records or assignment documents attached to this form in accordance with 37 CFR 3.73(c).

Please change the correspondence address for the application identified in the attached statement under 37 CFR 3.73(c) to:

☒ The address associated with Customer Number: **20350**

OR

☐ Firm or Individual Name			
Address			
City		State	Zip
Country			
Telephone		Email	

Assignee Name and Address:
The Chinese University of Hong Kong
Knowledge Transfer Office
Pi Ch'iu Building
Shatin, New Territories, Hong Kong SAR
People's Republic of China

A copy of this form, together with a statement under 37 CFR 3.73(c) (Form PTO/SB/96 or equivalent) is required to be filed in each application in which this form is used. The statement under 37 CFR 3.73(c) may be completed by one of the practitioners appointed in this form, and must identify the application in which this Power of Attorney is to be filed.

SIGNATURE of Assignee of Record
The individual whose signature and title is supplied below is authorized to act on behalf of the assignee

Signature		Date	26 Feb 2014
Name	[illegible] CHAN	Telephone	(852) 3943 1078
Title	Knowledge Transfer Office, The Chinese University of Hong Kong		

This collection of information is required by 37 CFR 1.31, 1.32 and 1.33. The information is required to obtain or retain a benefit by the public which is to file (and by the USPTO to process) an application. Confidentiality is governed by 35 U.S.C. 122 and 37 CFR 1.11 and 1.14. This collection is estimated to take 3 minutes to complete, including gathering, preparing, and submitting the completed application form to the USPTO. Time will vary depending upon the individual case. Any comments on the amount of time you require to complete this form and/or suggestions for reducing this burden, should be sent to the Chief Information Officer, U.S. Patent and Trademark Office, U.S. Department of Commerce, P.O. Box 1450, Alexandria, VA 22313-1450. DO NOT SEND FEES OR COMPLETED FORMS TO THIS ADDRESS. **SEND TO: Commissioner for Patents, P.O. Box 1450, Alexandria, VA 22313-1450.**
If you need assistance in completing the form, call 1-800-PTO-9199 and select option 2. #

Doc Code: TRACK1.REQ
Document Description: TrackOne Request

PTO/AIA/424 (04-14)

CERTIFICATION AND REQUEST FOR PRIORITIZED EXAMINATION UNDER 37 CFR 1.102(e) (Page 1 of 1)

First Named Inventor:	Yuk-Ming Dennis Lo	Nonprovisional Application Number (if known):	
Title of Invention:	DETERMINATION OF BASE MODIFICATIONS OF NUCLEIC ACIDS		

APPLICANT HEREBY CERTIFIES THE FOLLOWING AND REQUESTS PRIORITIZED EXAMINATION FOR THE ABOVE-IDENTIFIED APPLICATION.

1. The processing fee set forth in 37 CFR 1.17(i)(1) and the prioritized examination fee set forth in 37 CFR 1.17(c) have been filed with the request. The publication fee requirement is met because that fee, set forth in 37 CFR 1.18(d), is currently $0. The basic filing fee, search fee, and examination fee are filed with the request or have been already been paid. I understand that any required excess claims fees or application size fee must be paid for the application.

2. I understand that the application may not contain, or be amended to contain, more than four independent claims, more than thirty total claims, or any multiple dependent claims, and that any request for an extension of time will cause an outstanding Track I request to be dismissed.

3. The applicable box is checked below:

I. ☑ Original Application (Track One) - Prioritized Examination under § 1.102(e)(1)

i. (a) The application is an original nonprovisional utility application filed under 35 U.S.C. 111(a). This certification and request is being filed with the utility application via EFS-Web.

---OR---

(b) The application is an original nonprovisional plant application filed under 35 U.S.C. 111(a). This certification and request is being filed with the plant application in paper.

ii. An executed inventor's oath or declaration under 37 CFR 1.63 or 37 CFR 1.64 for each inventor, **or** the application data sheet meeting the conditions specified in 37 CFR 1.53(f)(3)(i) is filed with the application.

II. ☐ Request for Continued Examination - Prioritized Examination under § 1.102(e)(2)

i. A request for continued examination has been filed with, or prior to, this form.
ii. If the application is a utility application, this certification and request is being filed via EFS-Web.
iii. The application is an original nonprovisional utility application filed under 35 U.S.C. 111(a), or is a national stage entry under 35 U.S.C. 371.
iv. This certification and request is being filed prior to the mailing of a first Office action responsive to the request for continued examination.
v. No prior request for continued examination has been granted prioritized examination status under 37 CFR 1.102(e)(2).

Signature /David D. Hsu/	Date 8/17/2020
Name (Print/Typed) David D. Hsu	Practitioner Registration Number 71,659

Note: *This form must be signed in accordance with 37 CFR 1.33. See 37 CFR 1.4(d) for signature requirements and certifications. Submit multiple forms if more than one signature is required.**

☑ *Total of 1 forms are submitted.

Privacy Act Statement

The **Privacy Act of 1974 (P.L. 93-579)** requires that you be given certain information in connection with your submission of the attached form related to a patent application or patent. Accordingly, pursuant to the requirements of the Act, please be advised that: (1) the general authority for the collection of this information is 35 U.S.C. 2(b)(2); (2) furnishing of the information solicited is voluntary; and (3) the principal purpose for which the information is used by the U.S. Patent and Trademark Office is to process and/or examine your submission related to a patent application or patent. If you do not furnish the requested information, the U.S. Patent and Trademark Office may not be able to process and/or examine your submission, which may result in termination of proceedings or abandonment of the application or expiration of the patent.

The information provided by you in this form will be subject to the following routine uses:

1. The information on this form will be treated confidentially to the extent allowed under the Freedom of Information Act (5 U.S.C. 552) and the Privacy Act (5 U.S.C 552a). Records from this system of records may be disclosed to the Department of Justice to determine whether disclosure of these records is required by the Freedom of Information Act.
2. A record from this system of records may be disclosed, as a routine use, in the course of presenting evidence to a court, magistrate, or administrative tribunal, including disclosures to opposing counsel in the course of settlement negotiations.
3. A record in this system of records may be disclosed, as a routine use, to a Member of Congress submitting a request involving an individual, to whom the record pertains, when the individual has requested assistance from the Member with respect to the subject matter of the record.
4. A record in this system of records may be disclosed, as a routine use, to a contractor of the Agency having need for the information in order to perform a contract. Recipients of information shall be required to comply with the requirements of the Privacy Act of 1974, as amended, pursuant to 5 U.S.C. 552a(m).
5. A record related to an International Application filed under the Patent Cooperation Treaty in this system of records may be disclosed, as a routine use, to the International Bureau of the World Intellectual Property Organization, pursuant to the Patent Cooperation Treaty.
6. A record in this system of records may be disclosed, as a routine use, to another federal agency for purposes of National Security review (35 U.S.C. 181) and for review pursuant to the Atomic Energy Act (42 U.S.C. 218(c)).
7. A record from this system of records may be disclosed, as a routine use, to the Administrator, General Services, or his/her designee, during an inspection of records conducted by GSA as part of that agency's responsibility to recommend improvements in records management practices and programs, under authority of 44 U.S.C. 2904 and 2906. Such disclosure shall be made in accordance with the GSA regulations governing inspection of records for this purpose, and any other relevant (*i.e.*, GSA or Commerce) directive. Such disclosure shall not be used to make determinations about individuals.
8. A record from this system of records may be disclosed, as a routine use, to the public after either publication of the application pursuant to 35 U.S.C. 122(b) or issuance of a patent pursuant to 35 U.S.C. 151. Further, a record may be disclosed, subject to the limitations of 37 CFR 1.14, as a routine use, to the public if the record was filed in an application which became abandoned or in which the proceedings were terminated and which application is referenced by either a published application, an application open to public inspection or an issued patent.
9. A record from this system of records may be disclosed, as a routine use, to a Federal, State, or local law enforcement agency, if the USPTO becomes aware of a violation or potential violation of law or regulation.

I hereby certify that this correspondence is being filed via EFS-Web with the United States Patent and Trademark Office on August 17, 2020.

KILPATRICK TOWNSEND & STOCKTON LLP

By: /Heather A. Orosco/
Heather A. Orosco

PATENT
Attorney Docket No.: 080015-028410US-1199353
Client Ref. No.: 19/MED/901 US Reg

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

In re Application of:	Confirmation No.: Not Yet Known
Yuk Ming Dennis Lo et al.	Examiner: Not Yet Known
Application No.: Unassigned	Technology Center/Art Unit: Not Yet Known
Filed: Herewith	
For: DETERMINATION OF BASE MODIFICATIONS OF NUCLEIC ACIDS	PETITION TO ACCEPT COLOR DRAWINGS
Customer No.: 20350	

Commissioner for Patents
P.O. Box 1450
Alexandria, VA 22313-1450

Commissioner:

Applicant hereby petitions under 37 CFR § 1.84(a)(2) that the color drawing filed in the application be accepted. Applicant hereby petitions the Commissioner to accept the color drawings for FIGS. 1-16, 18A-26, 28-40, 42, 44, 47, 49-50, 52A, 54, 56A-58, 61, 63A-64B, 67A-B, 70A-B, 73-77, 79A-B, 79D, 80D, 82-85C, 88-94, 104, 108, 115, and 117 in connection with the above-identified application. The color drawings are believed to be necessary because the drawings include certain images rendered in color, from which information would be lost if the images were rendered as black-and-white line drawings.

The color drawings are concurrently being filed with this petition. The requisite fee for this Petition is being submitted herewith.

The Commissioner is authorized to charge any additional fees during pendency of this application, including any required extension of time fees, or credit any overpayment to Deposit Account Number 20-1430.

The following text comprises paragraph **[0013]**, the first paragraph under the heading "Brief Description of the Drawings":

The patent or application file contains at least one drawing executed in color. Copies of this patent or patent application publication with color drawing(s) will be provided by the Office upon request and payment of the necessary fee.

If the Examiner believes a telephone conference would expedite prosecution of this application, please contact the undersigned at (303) 571-4000 or dhsu@kilpatricktownsend.com.

Respectfully submitted,

/David D. Hsu/
David D. Hsu
Registration No. 71659

KILPATRICK TOWNSEND & STOCKTON LLP

DDH:hao

REMARKS/ARGUMENTS

Upon entry of the present amendment, claims 1-18, 70, and 71 will be pending in this application. Claims 18-69 have been canceled. No new matter is added.

CONCLUSION

In view of the foregoing, Applicant believes all claims now pending in this application are in condition for examination.

Except for the issue fees payable under 37 C.F.R. § 1.18, the Director is authorized to charge any additional fees during pendency of this application, including any required extension of time fees, or credit any overpayment to Deposit Account Number 20-1430. This paragraph is intended to be a constructive petition for extension of time in accordance with 37 C.F.R. § 1.136(a)(3).

If the Examiner believes a telephone conference would expedite prosecution of this application, please contact the undersigned at (303) 571-4000 or dhsu@kilpatricktownsend.com.

Respectfully submitted,

/David D. Hsu/
David D. Hsu
Registration No. 71659

KILPATRICK TOWNSEND & STOCKTON LLP

DDH:hao

AMENDMENTS TO THE CLAIMS

This listing of claims will replace all prior versions, and listings of claims in the application.

Listing of Claims:

1. (Original) A method for detecting a modification of a nucleotide in a nucleic acid molecule, the method comprising:

receiving an input data structure, the input data structure corresponding to a window of nucleotides sequenced in a sample nucleic acid molecule, wherein the sample nucleic acid molecule is sequenced by measuring pulses in an optical signal corresponding to the nucleotides, the input data structure comprising values for the following properties:

for each nucleotide within the window:

an identity of the nucleotide,

a position of the nucleotide with respect to a target position within the respective window,

a width of the pulse corresponding to the nucleotide, and

an interpulse duration representing a time between the pulse corresponding to the nucleotide and a pulse corresponding to a neighboring nucleotide;

inputting the input data structure into a model, the model trained by:

receiving a first plurality of first data structures, each first data structure of the first plurality of data structures corresponding to a respective window of nucleotides sequenced in a respective nucleic acid molecule of a plurality of first nucleic acid molecules, wherein each of the first nucleic acid molecules is sequenced by measuring pulses in the optical signal corresponding to the nucleotides, wherein the modification has a known first state in a nucleotide at a target position in each window of each first nucleic acid molecule, each first data structure comprising values for the same properties as the input data structure,

storing a plurality of first training samples, each including one of the first

plurality of first data structures and a first label indicating the first state of the nucleotide at the target position, and

optimizing, using the plurality of first training samples, parameters of the model based on outputs of the model matching or not matching corresponding labels of the first labels when the first plurality of first data structures is input to the model, wherein an output of the model specifies whether the nucleotide at the target position in the respective window has the modification,

determining, using the model, whether the modification is present in a nucleotide at the target position within the window in the input data structure.

2. (Original) The method of claim 1, wherein:

the input data structure is one input data structure of a plurality of input data structures,

the sample nucleic acid molecule is one sample nucleic acid molecule of a plurality of sample nucleic acid molecules,

the plurality of sample nucleic acid molecules are obtained from a biological sample of a subject, and

each input data structure corresponds to a respective window of nucleotides sequenced in a respective sample nucleic acid molecule of the plurality of sample nucleic acid molecules, and

the method further comprising:

receiving the plurality of input data structures,

inputting the plurality of input data structures into the model, and

determining, using the model, whether a modification is present in a nucleotide at a target location in the respective window of each input data structure.

3. (Original) The method of claim 2, further comprising:

determining the modification is present at one or more nucleotides, and

determining a classification of a disorder using the presence of the modification at one or more nucleotides.

4. (Original) The method of claim 3, wherein the disorder comprises cancer.

5. (Original) The method of claim 3, further comprising:

determining that the classification of the disorder is that the subject has the disorder, and

treating the subject for the disorder.

6. (Original) The method of claim 3, wherein determining the classification of the disorder uses the number of modifications or the sites of the modifications.

7. (Original) The method of claim 2, further comprising:

determining the modification is present at one or more nucleotides, and

determining a clinically-relevant DNA fraction, a fetal methylation profile, a maternal methylation profile, a presence of an imprinting gene region, or a tissue of origin using the presence of the modification at one or more nucleotides.

8. (Original) The method of claim 2, wherein each sample nucleic acid molecule of the plurality of sample nucleic acid molecules has a size greater than a cutoff size.

9. (Original) The method of claim 2, wherein:

the plurality of sample nucleic acid molecules align to a plurality of genomic regions,

for each genomic region of the plurality of genomic regions:

a number of sample nucleic acid molecules is aligned to the genomic region,

the number of sample nucleic acid molecules is greater than a cutoff number.

10. (Original) The method of claim 1, further comprising sequencing the sample nucleic acid molecule.

11. (Original) The method of claim 1, wherein the model includes a machine learning model, a principal component analysis, a convolutional neural network, or a logistic

regression.

12. (Original) The method of claim 1, wherein:

the window of nucleotides corresponding to the input data structure comprises nucleotides on a first strand of the sample nucleic acid molecule and nucleotides on a second strand of the sample nucleic acid molecule, and

the input data structure further comprises for each nucleotide within the window a value of a strand property, the strand property indicating the nucleotide being present on either the first strand or the second strand.

13. (Original) The method of claim 12, wherein the sample nucleic acid molecule is a circular DNA molecule formed by:

cutting a double-stranded DNA molecule using a Cas9 complex to form a cut double-stranded DNA molecule, and

ligating a hairpin adaptor onto an end of the cut double-stranded DNA molecule.

14. (Original) The method of claim 1, wherein the nucleotides within the window are determined using a circular consensus sequence and without alignment of the sequenced nucleotides to a reference genome.

15. (Original) The method of claim 1, wherein each nucleotide within the window is enriched or filtered.

16. (Original) The method of claim 15, wherein each nucleotide within the window is enriched by:

cutting a double-stranded DNA molecule using a Cas9 complex to form a cut double-stranded DNA molecule, and ligating a hairpin adaptor onto an end of the cut double-stranded DNA molecule, or

filtered by:

selecting double-stranded DNA molecules having a size with a size range.

17. (Original) The method of claim 1, wherein nucleotides within the window are determined without using a circular consensus sequence and without alignment of the

sequenced nucleotides to a reference genome.

18-69. (Canceled)

70. (Original) The method of claim 1, wherein the optical signal is a fluorescence signal from a dye-labeled nucleotide.

71. (Original) The method of claim 1, wherein each window associated with the first plurality of data structures comprises 4 consecutive nucleotides on a first strand of each first nucleic acid molecule.

I hereby certify that this correspondence is being filed via EFS-Web with the United States Patent and Trademark Office on August 17, 2020.

KILPATRICK TOWNSEND & STOCKTON LLP

By: /Heather A. Orosco/
Heather A. Orosco

PATENT
Attorney Docket No.: 080015-028410US-1199353
Client Ref. No.: 19/MED/901 US Reg

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

In re Application of:

Yuk-Ming Dennis Lo et al.

Application No.: Unassigned

Filed: Herewith

For: DETERMINATION OF BASE MODIFICATIONS OF NUCLEIC ACIDS

Customer No.: 20350

Confirmation No.: Not Yet Known

Examiner: Not Yet Known

Technology Center/Art Unit: Not Yet Known

PRELIMINARY AMENDMENT

Mail Stop Amendment
Commissioner for Patents
P.O. Box 1450
Alexandria, VA 22313-1450

Commissioner:

Prior to examination of the above-referenced application, please enter the following amendments and consider the following remarks:

Amendments to the Claims are reflected in the listing of claims which begins on page 2 of this paper.

Remarks/Arguments begin on page 7 of this paper.

ABSTRACT

Systems and methods for using determination of base modification in analyzing nucleic acid molecules and acquiring data for analysis of nucleic acid molecules are described herein. Base modifications may include methylations. Methods to determine base modifications may include using features derived from sequencing. These features may include the pulse width of an optical signal from sequencing bases, the interpulse duration of bases, and the identity of the bases. Machine learning models can be trained to detect the base modifications using these features. The relative modification or methylation levels between haplotypes may indicate a disorder. Modification or methylation statuses may also be used to detect chimeric molecules.

WHAT IS CLAIMED IS:

1. A method for detecting a modification of a nucleotide in a nucleic acid molecule, the method comprising:

receiving an input data structure, the input data structure corresponding to a window of nucleotides sequenced in a sample nucleic acid molecule, wherein the sample nucleic acid molecule is sequenced by measuring pulses in an optical signal corresponding to the nucleotides, the input data structure comprising values for the following properties:

for each nucleotide within the window:

an identity of the nucleotide,

a position of the nucleotide with respect to a target position within the respective window,

a width of the pulse corresponding to the nucleotide, and

an interpulse duration representing a time between the pulse corresponding to the nucleotide and a pulse corresponding to a neighboring nucleotide;

inputting the input data structure into a model, the model trained by:

receiving a first plurality of first data structures, each first data structure of the first plurality of data structures corresponding to a respective window of nucleotides sequenced in a respective nucleic acid molecule of a plurality of first nucleic acid molecules, wherein each of the first nucleic acid molecules is sequenced by measuring pulses in the optical signal corresponding to the nucleotides, wherein the modification has a known first state in a nucleotide at a target position in each window of each first nucleic acid molecule, each first data structure comprising values for the same properties as the input data structure,

storing a plurality of first training samples, each including one of the first plurality of first data structures and a first label indicating the first state of the nucleotide at the target position, and

optimizing, using the plurality of first training samples, parameters of the model based on outputs of the model matching or not matching corresponding labels of the first labels when the first plurality of first data structures is input to the model,

wherein an output of the model specifies whether the nucleotide at the target position in the respective window has the modification,

determining, using the model, whether the modification is present in a nucleotide at the target position within the window in the input data structure.

2. The method of claim 1, wherein:

the input data structure is one input data structure of a plurality of input data structures,

the sample nucleic acid molecule is one sample nucleic acid molecule of a plurality of sample nucleic acid molecules,

the plurality of sample nucleic acid molecules are obtained from a biological sample of a subject, and

each input data structure corresponds to a respective window of nucleotides sequenced in a respective sample nucleic acid molecule of the plurality of sample nucleic acid molecules, and

the method further comprising:

receiving the plurality of input data structures,

inputting the plurality of input data structures into the model, and

determining, using the model, whether a modification is present in a nucleotide at a target location in the respective window of each input data structure.

3. The method of claim 2, further comprising:

determining the modification is present at one or more nucleotides, and

determining a classification of a disorder using the presence of the modification at one or more nucleotides.

4. The method of claim 3, wherein the disorder comprises cancer.

5. The method of claim 3, further comprising:

determining that the classification of the disorder is that the subject has the disorder, and

treating the subject for the disorder.

6. The method of claim 3, wherein determining the classification of the disorder uses the number of modifications or the sites of the modifications.

7. The method of claim 2, further comprising:

determining the modification is present at one or more nucleotides, and

determining a clinically-relevant DNA fraction, a fetal methylation profile, a maternal methylation profile, a presence of an imprinting gene region, or a tissue of origin using the presence of the modification at one or more nucleotides.

8. The method of claim 2, wherein each sample nucleic acid molecule of the plurality of sample nucleic acid molecules has a size greater than a cutoff size.

9. The method of claim 2, wherein:

the plurality of sample nucleic acid molecules align to a plurality of genomic regions,

for each genomic region of the plurality of genomic regions:

a number of sample nucleic acid molecules is aligned to the genomic region,

the number of sample nucleic acid molecules is greater than a cutoff number.

10. The method of claim 1, further comprising sequencing the sample nucleic acid molecule.

11. The method of claim 1, wherein the model includes a machine learning model, a principal component analysis, a convolutional neural network, or a logistic regression.

12. The method of claim 1, wherein:

the window of nucleotides corresponding to the input data structure comprises nucleotides on a first strand of the sample nucleic acid molecule and nucleotides on a second strand of the sample nucleic acid molecule, and

the input data structure further comprises for each nucleotide within the window a value of a strand property, the strand property indicating the nucleotide being present on either the first strand or the second strand.

13. The method of claim 12, wherein the sample nucleic acid molecule is a circular DNA molecule formed by:

cutting a double-stranded DNA molecule using a Cas9 complex to form a cut double-stranded DNA molecule, and

ligating a hairpin adaptor onto an end of the cut double-stranded DNA molecule.

14. The method of claim 1, wherein the nucleotides within the window are determined using a circular consensus sequence and without alignment of the sequenced nucleotides to a reference genome.

15. The method of claim 1, wherein each nucleotide within the window is enriched or filtered.

16. The method of claim 15, wherein each nucleotide within the window is enriched by:

cutting a double-stranded DNA molecule using a Cas9 complex to form a cut double-stranded DNA molecule, and ligating a hairpin adaptor onto an end of the cut double-stranded DNA molecule, or

filtered by:

selecting double-stranded DNA molecules having a size with a size range.

17. The method of claim 1, wherein nucleotides within the window are determined without using a circular consensus sequence and without alignment of the sequenced nucleotides to a reference genome.

18. A method for detection of a modification of a nucleotide in a nucleic acid molecule, the method comprising:

receiving a first plurality of first data structures, each first data structure of the first plurality of first data structures corresponding to a respective window of nucleotides sequenced in a respective nucleic acid molecule of a plurality of first nucleic acid molecules, wherein each of the first nucleic acid molecules is sequenced by measuring pulses in an optical signal corresponding to the nucleotides, wherein the modification has a known first state in the

nucleotide at a target position in each window of each first nucleic acid molecule, each first data structure comprising values for the following properties:

for each nucleotide within the window:

an identity of the nucleotide,

a position of the nucleotide with respect to the target position within the respective window,

a width of the pulse corresponding to the nucleotide, and

an interpulse duration representing a time between the pulse corresponding to the nucleotide and a pulse corresponding to a neighboring nucleotide,

storing a plurality of first training samples, each including one of the first plurality of first data structures and a first label indicating the first state for the modification of the nucleotide at the target position; and

training a model using the plurality of first training samples by optimizing parameters of the model based on outputs of the model matching or not matching corresponding labels of the first labels when the first plurality of first data structures is input to the model, wherein an output of the model specifies whether the nucleotide at the target position in the respective window has the modification.

19. The method of claim 18, further comprising:

receiving a second plurality of second data structures, each second data structure of the second plurality of second data structures corresponding to a respective window of nucleotides sequenced in a respective nucleic acid molecule of a plurality of second nucleic acid molecules, wherein the modification has a known second state in a nucleotide at a target position within each window of each second nucleic acid molecule, each second data structure comprising values for the same properties as the first plurality of first data structures;

storing a plurality of second training samples, each including one of the second plurality of second data structures and a second label indicating the second state of the nucleotide at the target position;

wherein training:

the first state or the second state is that the modification is present and the other state is that the modification is absent,

the model further comprises using the plurality of second training samples by optimizing parameters of the model based on outputs of the model matching or not matching corresponding labels of the second labels when the second plurality of second data structures are input to the model.

20. The method of claim 19, wherein the plurality of first nucleic acid molecules is the same as the plurality of the second nucleic acid molecules.

21. The method of claim 19, wherein the modification comprises methylation, wherein the plurality of first nucleic acid molecules is generated using multiple displacement amplification with methylated nucleotides of a first type, and wherein the plurality of second nucleic acid molecules is generated using multiple displacement amplification with unmethylated nucleotides of the first type.

22. The method of claim 18, wherein the optical signal is a fluorescence signal from a dye-labeled nucleotide.

23. The method of claim 18, wherein each window associated with the first plurality of data structures comprises 4 consecutive nucleotides on a first strand of each first nucleic acid molecule.

24. The method of claim 23, wherein the windows associated with the first plurality of data structures comprise the same number of consecutive nucleotides.

25. The method of claim 18, wherein:

each window associated with the first plurality of data structures comprises nucleotides on a first strand of the first nucleic acid molecule and nucleotides on a second strand of the first nucleic acid molecule, and

each first data structure further comprises for each nucleotide within the window a value of a strand property, the strand property indicating the nucleotide being present on either the first strand or the second strand.

26. The method of claim 18, wherein the neighboring nucleotide is the adjacent nucleotide.

27. The method of claim 18, wherein the width of the pulse is the width of the pulse at half the maximum value of the pulse.

28. The method of claim 18, wherein the interpulse duration is the time between the maximum value of the pulse associated with the nucleotide and the maximum value of the pulse associated with the neighboring nucleotide.

29. The method of claim 18, wherein the model comprises a convolutional neural network comprising:

a set of convolutional filters configured to filter the first plurality of data structures,

an input layer configured to receive the filtered first plurality of data structures,

a plurality of hidden layers including a plurality of nodes, a first layer of the plurality of hidden layers coupled to the input layer; and

an output layer coupled to a last layer of the plurality of hidden layers and configured to output an output data structure, the output data structure comprising the properties.

30. The method of claim 18, wherein the modification comprises a methylation of the nucleotide at the target position.

31. The method of claim 30, wherein the known first states include a methylated state for a first portion of the first data structures and an unmethylated state for a second portion of the first data structures.

32. The method of claim 30, wherein the methylation comprises 4mC (N4-methylcytosine), 5mC (5-methylcytosine), 5hmC (5-hydroxymethylcytosine), 5fC (5-formylcytosine), 5caC (5-carboxylcytosine), 1mA (N1-methyladenine), 3mA (N3-methyladenine), 6mA (N6-methyladenine), 7mA (N7-methyladenine), 3mC (N3-methylcytosine), 2mG (N2-methylguanine), 6mG (O6-methylguanine), 7mG (N7-methylguanine), 3mT (N3-methylthymine), or 4mT (O4-methylthymine).

33. The method of claim 18, wherein the modification comprises an oxidation change.

34. The method of claim 18, wherein each data structure further comprises a value for a height of the pulse corresponding to each nucleotide within the window.

35. The method of claim 18, wherein the optical signal corresponding to the nucleotides results from the nucleotides or a tag associated with the nucleotides.

36. The method of claim 18, wherein each target position is the center of the respective window.

37. The method of claim 18, wherein the modification is absent in each window of each first nucleic acid molecule.

38. The method of claim 18, wherein:

each first data structure of the plurality of first data structures excludes first nucleic acid molecules with an interpulse duration or a width of a pulse below a cutoff value.

39. The method of claim 18, wherein:

the modification comprises a methylation, and

the plurality of first training samples is generated by:

amplifying a plurality of nucleic acid molecules using a set of nucleotides, wherein the set of nucleotides includes 6mA at a specified ratio.

40. The method of claim 39, wherein the methylation comprises 6mA (N6-methyladenine).

41. The method of claim 1 or claim 18, wherein at least some of the plurality of first nucleic acid molecules each include a first portion corresponding to a first reference sequence and a second portion corresponding to a second reference sequence that is disjoint from the first reference sequence.

42. The method of claim 1 or claim 18, further comprising:

validating the model using a plurality of chimeric nucleic acid molecules, each including a first portion corresponding to a first reference sequence and a second portion

corresponding to a second reference sequence, wherein the first portion has a first methylation pattern and the second portion has a second methylation pattern.

43. The method of claim 41 or claim 42, wherein the first portion is treated with a methylase.

44. The method of claim 43, wherein the second portion corresponds to an unmethylated portion of the second reference sequence.

45. The method of claim 41 or claim 42, wherein the first reference sequence is human, and wherein the second reference sequence is from a different animal.

46. A method of analyzing a biological sample of an organism having a first haplotype and a second haplotype in a first chromosomal region, the biological sample including DNA molecules, the method comprising:

analyzing a plurality of DNA molecules from the biological sample, wherein analyzing a DNA molecule includes:

identifying a location of the DNA molecule in a reference human genome;

determining a respective allele of the DNA molecule; and

determining whether the DNA molecule is methylated at one or more genomic sites;

identifying one or more heterozygous loci of a first portion of the first chromosomal region, each heterozygous locus including a corresponding first allele in the first haplotype and a corresponding second allele in the second haplotype;

identifying a first set of the plurality of DNA molecules that each:

is located at any one of the one or more heterozygous loci,

includes the corresponding first allele of the heterozygous locus, and

includes at least one of N genomic sites, N being an integer greater than or equal to one;

determining a first methylation level of the first portion of the first haplotype using the first set of the plurality of DNA molecules;

identifying a second set of the plurality of DNA molecules that each:

is located at any one of the one or more heterozygous loci,

includes the corresponding second allele of the heterozygous locus, and includes at least one of the N genomic sites;

determining a second methylation level of the first portion of the second haplotype using the second set of the plurality of DNA molecules;

calculating a value of a parameter using the first methylation level and the second methylation level;

comparing the value of the parameter to a reference value; and

determining a classification of a disorder in the organism using the comparison of the value of the parameter to the reference value.

47. The method of claim 46, wherein the first methylation level is determined using single strand methylation levels of the first set of the plurality of DNA molecules, and wherein the second methylation level is determined using single strand methylation levels of the second set of the plurality of DNA molecules.

48. The method of claim 46, wherein the first methylation level is determined using single molecule, double-stranded DNA methylation levels of the first set of the plurality of DNA molecules, and wherein the second methylation level is determined using single molecule, double-stranded DNA methylation of the second set of the plurality of DNA molecules.

49. The method of claim 46, wherein the disorder is cancer.

50. The method of claim 46, wherein the parameter is a separation value.

51. The method of claim 46, further comprising:

determining a plurality of first methylation levels for a plurality of portions of the first haplotype,

determining a plurality of second methylation levels for a plurality of portions of the second haplotype, each portion of plurality of portions of the second haplotype being complementary to a portion of the plurality of portions of the first haplotype,

for each portion of the plurality of portions of the second haplotype:

calculating a separation value using the second methylation level of the portion of the second haplotype and the first methylation level of the complementary portion of the first haplotype, and

comparing the separation value to a cutoff value,

wherein:

the first portion of the first haplotype is complementary to the first portion of the second haplotype, and

the parameter includes a number of portions of the second haplotype where the separation value exceeds the cutoff value.

52. The method of claim 51, wherein the cutoff value is determined from tissues not having the disorder.

53. The method of claim 51, wherein each portion of the plurality of portions of the first haplotype has a length of greater than or equal to 5 kb.

54. The method of claim 46, further comprising:

determining a plurality of first methylation levels for a plurality of portions of the first haplotype,

determining a plurality of second methylation levels for a plurality of portions of the second haplotype, each portion of the plurality of portions of the second haplotype being complementary to a portion of the plurality of portions of the first haplotype,

for each portion of the plurality of portions of the second haplotype:

calculating a separation value using the second methylation level of the portion of the second haplotype and the first methylation level of the complementary portion of the first haplotype,

wherein:

the first portion of the first haplotype is complementary to the first portion of the second haplotype, and

the parameter includes a sum of the separation values.

55. The method of claim 46, further comprising:

determining a plurality of first methylation levels for a plurality of portions of the first haplotype,

determining a plurality of second methylation levels for a plurality of portions of the second haplotype, each portion of the plurality of portions of the second haplotype being complementary to a portion of the plurality of portions of the first haplotype,

for each portion of the plurality of portions of the second haplotype:

calculating a separation value using the second methylation level of the portion of the second haplotype and the first methylation level of the complementary portion of the first haplotype, and

comparing the separation value to a cutoff value to identify whether the portion has aberrant separation between the first methylation level and the second methylation level,

wherein determining the classification of the disorder in the organism includes comparing a pattern of portions having an aberrant separation to a reference pattern.

56. The method of claim 46, wherein the classification of the disorder is a likelihood of the disorder.

57. The method of claim 46, wherein:

the first portion of the first haplotype and the first portion of the second haplotype form a circular DNA molecule, and

determining the first methylation level of the first portion of the first haplotype comprises using data from the circular DNA molecule.

58. The method of claim 57, wherein the circular DNA molecule is formed by:

cutting a double-stranded DNA molecule using a Cas9 complex to form a cut double-stranded DNA molecule, and

ligating a hairpin adaptor onto an end of the cut double-stranded DNA molecule.

59. The method of claim 46, wherein:

the first portion of the first haplotype is longer than or equal to 1 kb.

60. The method of claim 46, wherein the reference value is determined using a reference tissue without the disorder.

61. The method of claim 46, wherein the disorder is an imprinting disorder.

62. A method of detecting chimeric molecules in a biological sample, the method comprising:

for each of a plurality of DNA molecules from the biological sample:

performing single molecule sequencing of the DNA molecule to obtain a sequence read that provides a methylation status at each of N sites, N being 5 or more, wherein the methylation statuses of the sequence read form a methylation pattern;

sliding the methylation pattern over one or more reference patterns that correspond to chimeric molecules that have two portions from two parts of a reference human genome, the one or more reference patterns including a change between methylated statuses and unmethylated statuses; and

identifying a match position between the methylation pattern and a first reference pattern of the one or more reference patterns, the match position identifying a junction between the two parts of the reference human genome in the sequence read; and

outputting the junction as a location of a gene fusion in a chimeric molecule.

63. The method of claim 62, wherein the match position is output to an alignment function, the method further comprising:

refining the location of the gene fusion by:

aligning a first portion of the sequence read to a first part of the reference human genome, the first portion being before the junction; and

aligning a second portion of the sequence read to a second part of the reference human genome, the second portion being after the junction, wherein the first part of the reference human genome is at least 1 kb apart from the second part of the human reference genome.

64. The method of claim 62, further comprising comparing the junctions of the chimeric molecules to each other to confirm the location of the gene fusion.

65. A computer product comprising a non-transitory computer readable medium storing a plurality of instructions that when executed control a computer system to perform the method of any one of the preceding claims.

66. A system comprising:

the computer product of claim 65; and

one or more processors for executing instructions stored on the computer readable medium.

67. A system comprising means for performing any of the above methods.

68. A system comprising one or more processors configured to perform any of the above methods.

69. A system comprising modules that respectively perform the steps of any of the above methods.

70. The method of claim 1, wherein the optical signal is a fluorescence signal from a dye-labeled nucleotide.

71. The method of claim 1, wherein each window associated with the first plurality of data structures comprises 4 consecutive nucleotides on a first strand of each first nucleic acid molecule.

PATENT APPLICATION

DETERMINATION OF BASE MODIFICATIONS OF NUCLEIC ACIDS

Inventors:

Yuk-Ming Dennis Lo, a citizen of United Kingdom, residing at
Homantin, Kowloon, Hong Kong SAR China

Rossa Wai Kwun Chiu, a citizen of Australia, residing at
Shatin, New Territories, Hong Kong SAR China

Kwan Chee Chan, a citizen of Hong Kong, residing at
Jordan, Hong Kong SAR China

Peiyong Jiang, a citizen of Hong Kong, residing at
Shatin, New Territories, Hong Kong SAR China

Suk Hang Cheng, a citizen of Hong Kong, residing at
Fanling, New Territories, Hong Kong SAR China

Wenlei Peng, a citizen of the People's Republic of China, residing at
Shatin, New Territories, Hong Kong SAR China

On Yee Tse, a citizen of Hong Kong, residing at
Fanling, New Territories, Hong Kong SAR China

Assignee: The Chinese University of Hong Kong
Office of Research and Knowledge Transfer Services
Room 301, Pi Ch'iu Building
Shatin, New Territories Hong Kong

Entity: Small

DETERMINATION OF BASE MODIFICATIONS OF NUCLEIC ACIDS

CROSS-REFERENCES TO RELATED APPLICATIONS

[0001] The present application claims the benefit of priority to U.S. Provisional Application No. 63/051,210, entitled "DETERMINATION OF BASE MODIFICATIONS OF NUCLEIC ACIDS," filed on July 13, 2020; U.S. Provisional Application No. 63/019,790, entitled "DETERMINATION OF BASE MODIFICATIONS OF NUCLEIC ACIDS," filed on May 4, 2020; U.S. Provisional Application No. 62/991,891, entitled "DETERMINATION OF BASE MODIFICATIONS OF NUCLEIC ACIDS," filed on March 19, 2020; U.S. Provisional Application No. 62/970,586, entitled "DETERMINATION OF BASE MODIFICATIONS OF NUCLEIC ACIDS," filed on February 5, 2020; and U.S. Provisional Application No. 62/887,987, entitled "DETERMINATION OF BASE MODIFICATIONS OF NUCLEIC ACIDS," filed on August 16, 2019, the entire contents of all of which are herein incorporated by reference for all purposes.

BACKGROUND

[0002] The existence of base modifications in nucleic acids varies throughout different organisms including viruses, bacteria, plants, fungi, nematodes, insects, and vertebrates (e.g. humans), etc. The most common base modifications are the addition of a methyl group to different DNA bases at different positions, so-called methylation. Methylation has been found on cytosines, adenines, thymines and guanines, such as 5mC (5-methylcytosine), 4mC (N4-methylcytosine), 5hmC (5-hydroxymethylcytosine), 5fC (5-formylcytosine), 5caC (5-carboxylcytosine), 1mA (N1-methyladenine), 3mA (N3-methyladenine), 7mA (N7-methyladenine), 3mC (N3-methylcytosine), 2mG (N2-methylguanine), 6mG (O6-methylguanine), 7mG (N7-methylguanine), 3mT (N3-methylthymine), and 4mT (O4-methylthymine). In vertebrate genomes, 5mC is the most common type of base methylation, followed by that for guanine (i.e. in the CpG context).

[0003] DNA methylation is essential for mammalian development and has notable roles in gene expression and silencing, embryonic development, transcription, chromatin structure, X chromosome inactivation, protection against activity of the repetitive elements, maintenance of genomic stability during mitosis, and the regulation of parent-of-origin genomic imprinting.

[0004] DNA methylation plays many important roles in the silencing of promoters and enhancers in a coordinated manner (Robertson, 2005; Smith and Meissner, 2013). Many human diseases have been found to be associated with aberrations of DNA methylation, including but not limited to the process of carcinogenesis, imprinting disorders (e.g. Beckwith–Wiedemann syndrome and Prader–Willi syndrome), repeat-instability diseases (e.g. fragile X syndrome), autoimmune disorders (e.g. systemic lupus erythematosus), metabolic disorders (e.g. type I and type II diabetes), neurological disorders, aging, etc.

[0005] The accurate measurement of methylomic modification on DNA molecules would have numerous clinical implications. One widely used method to measure DNA methylation is through the use of bisulfite sequencing (BS-seq) (Lister *et al.*, 2009; Frommer *et al.*, 1992). In this approach, DNA samples are first treated with bisulfite which converts unmethylated cytosine (i.e. C) to uracil. In contrast, the methylated cytosine remains unchanged. The bisulfite modified DNA is then analyzed by DNA sequencing. In another approach, following bisulfite conversion, the modified DNA is then subjected to polymerase chain reaction (PCR) amplification using primers that can differentiate bisulfite converted DNA of different methylation profiles (Herman *et al.*, 1996). This latter approach is called methylation-specific PCR.

[0006] One disadvantage of such bisulfite-based approaches is that the bisulfite conversion step has been reported to significantly degrade the majority of the treated DNA (Grunau, 2001). Another disadvantage is that the bisulfite conversion step would create strong CG biases (Olova *et al.*, 2018), resulting in the reduction of signal-to-noise ratios typically for DNA mixtures with heterogeneous methylation states. Furthermore, bisulfite sequencing would not be able to sequence long DNA molecules because of the degradation of DNA during bisulfite treatment. Thus, there is a need to determine the modification of bases of nucleic acids, without prior chemical (e.g. bisulfite conversion) and nucleic acid amplification (e.g. using the PCR).

BRIEF SUMMARY

[0007] We have developed a new method that, in one embodiment, allows the determination of base modifications, such as 5mC in nucleic acids without template DNA pre-treatment such as enzymatic and/or chemical conversions, or protein and/or antibody binding. While such template DNA pre-treatment is not necessary for the determination of the base modifications, in examples that are shown, certain pre-treatment (e.g. digestion with restriction enzymes) may serve to enhance aspects of the invention (e.g. allowing the enrichment of CpG sites for analysis). The embodiments present in this disclosure could be used for detecting different types of base modification, for example, including but not limited to 4mC, 5hmC, 5fC, and 5caC, 1mA, 3mA, 7mA, 3mC, 2mG, 6mG, 7mG, 3mT, and 4mT, etc. Such embodiments can make use of features derived from sequencing, such as kinetic features, that are affected by the various base modifications, as well as an identity of nucleotides in a window around a target position whose methylation status is determined.

[0008] Embodiments of the present invention can be used for, but is not limited to, single molecule sequencing. One type of single molecule sequencing is single molecule, real-time sequencing in which the progress of the sequencing of a single DNA molecule is monitored in real-time. One type of single molecule, real-time sequencing is that commercialized by Pacific Biosciences using their Single Molecule, Real-Time (SMRT) system. Methods may use the pulse width of a signal from sequencing bases, the interpulse duration (IPD) of bases, and the identity of the bases in order to detect a modification in a base or in a neighboring base. Another single molecule system is that based on nanopore sequencing. One example of a nanopore sequencing system is that commercialized by Oxford Nanopore Technologies.

[0009] The methods we have developed can serve as tools to detect base modifications in biological samples to assess the methylation profiles in the samples for various purposes including but not limited to research and diagnostic purposes. The detected methylation profiles can be used for different analysis. The methylation profiles can be used to detect the origin of DNA (e.g., maternal or fetal, tissue, bacterial, or DNA obtained from tumor cells enriched from the blood of a cancer patient). Detection of aberrant methylation profiles in tissues aids the

identification of developmental disorders in individuals, identify and prognosticate tumors or malignancies.

[0010] Embodiments of the present invention may include analyzing the relative methylation levels of haplotypes of an organism. An imbalance in the methylation levels between the two haplotypes may be used to determine a classification of a disorder. A higher imbalance may indicate the presence of a disorder or a more severe disorder. The disorder may include cancer.

[0011] Methylation patterns in a single molecule can identify chimera and hybrid DNA. Chimeric and hybrid molecules may include sequences from two different genes, chromosomes, organelles (e.g. mitochondria, nucleus, chloroplasts), organisms (mammals, bacteria, viruses, etc.), and/or species. Detecting junctions of chimeric or hybrid DNA molecules may allow for detecting gene fusions for various disorders or diseases, including cancer, prenatal, or congenital disorders.

[0012] A better understanding of the nature and advantages of embodiments of the present invention may be gained with reference to the following detailed description and the accompanying drawings.

BRIEF DESCRIPTION OF THE DRAWINGS

[0013] The patent or application file contains at least one drawing executed in color. Copies of this patent or patent application publication with color drawing(s) will be provided by the Office upon request and payment of the necessary fee.

[0014] **FIG. 1** illustrates SMRT sequencing of molecules carrying base modifications according to embodiments of the present invention.

[0015] **FIG. 2** illustrates SMRT sequencing of molecules carrying methylated and unmethylated CpG sites according to embodiments of the present invention.

[0016] **FIG. 3** illustrates interpulse durations and pulse width according to embodiments of the present invention.

[0017] **FIG. 4** shows an example of a measurement window of the Watson strand of DNA for detecting a base modification according to embodiments of the present invention.

[0018] **FIG. 5** shows an example of a measurement window of the Crick strand of DNA for detecting a base modification according to embodiments of the present invention.

[0019] **FIG. 6** shows an example of a measurement window by combining data from the Watson strand of DNA and its complementary Crick strand for detecting any base modification according to embodiments of the present invention.

[0020] **FIG. 7** shows an example of a measurement window by combining data from the Watson strand of DNA and its nearby region's Crick strand for detecting any base modification according to embodiments of the present invention.

[0021] **FIG. 8** shows examples of measurement windows of the Watson strand, the Crick strand, and both strands for determining methylation states at CpG sites according to embodiments of the present invention.

[0022] **FIG. 9** shows a general procedure of constructing analytical, computational, mathematical, or statistical models for classifying base modifications according to embodiments of the present invention.

[0023] **FIG. 10** shows a general procedure of classifying base modifications according to embodiments of the present invention.

[0024] **FIG. 11** shows a general procedure of constructing the analytical, computational, mathematical, or statistical models for classifying methylation states at CpG sites using samples with known methylation states of the Watson strand according to embodiments of the present invention.

[0025] **FIG. 12** shows a general procedure of classifying methylation states of the Watson strand for an unknown sample according to embodiments of the present invention.

[0026] **FIG. 13** shows a general procedure of constructing the analytical, computational, mathematical, or statistical models for classifying methylation states at CpG sites using samples

with known methylation states of the Crick strand according to embodiments of the present invention.

[0027] **FIG. 14** shows a general procedure of classifying methylation states of the Crick strand for an unknown sample according to embodiments of the present invention.

[0028] **FIG. 15** shows a general procedure of constructing the statistical models for classifying methylation states at CpG sites using samples with known methylation states from both the Watson and Crick strands according to embodiments of the present invention.

[0029] **FIG. 16** shows a general procedure of classifying methylation states of an unknown sample from the Watson and Crick strands according to embodiments of the present invention.

[0030] **FIGS. 17A** and **17B** show the performance of a training dataset and a testing dataset for determining methylation according to embodiments of the present invention.

[0031] **FIGS. 18A** and **18B** show the performance of a training dataset and a testing dataset for determining methylation according to embodiments of the present invention.

[0032] **FIGS. 19A** and **19B** show the performance of a training dataset and a testing dataset at different sequencing depths for determining methylation according to embodiments of the present invention.

[0033] **FIGS. 20A** and **20B** show the performance of a training dataset and a testing dataset for different strands for determining methylation according to embodiments of the present invention.

[0034] **FIGS. 21A** and **21B** show the performance of a training dataset and a testing dataset for different measurement windows for determining methylation according to embodiments of the present invention.

[0035] **FIGS. 22A** and **22B** show the performance of a training dataset and a testing dataset for different measurement windows using downstream bases only for determining methylation according to embodiments of the present invention.

[0036] **FIGS. 23A** and **23B** show the performance of a training dataset and a testing dataset for different measurement windows using upstream bases only for determining methylation according to embodiments of the present invention.

[0037] **FIG. 24** shows the performance of methylation analysis using kinetic patterns associated with downstream and upstream bases using asymmetric flanking sizes in the training dataset according to embodiments of the present invention.

[0038] **FIG. 25** shows the performance of methylation analysis using kinetic patterns associated with downstream and upstream bases using asymmetric flanking sizes in the testing dataset according to embodiments of the present invention.

[0039] **FIG. 26** shows the relative importance of features in regard to classification of methylation states at CpG sites according to embodiments of the present invention.

[0040] **FIG. 27** shows the performance of the motif-based IPD analysis for methylation detection without using the pulse width signal according to embodiments of the present invention.

[0041] **FIG. 28** is a graph of a principal component analysis technique using 2-nt upstream and 6-nt downstream of a cytosine that is subjected to methylation analysis according to embodiments of the present invention.

[0042] **FIG. 29** is a graph of a performance comparison between a method using principal component analysis and a method using a convolutional neural network according to embodiments of the present invention.

[0043] **FIGS. 30A** and **30B** show the performance of a training dataset and a testing dataset for different analytical, computational, mathematical, or statistical models using upstream bases only for determining methylation according to embodiments of the present invention.

[0044] **FIG. 31A** shows an example of one approach for generating molecules with unmethylated adenines by whole genome amplification according to embodiments of the present invention.

[0045] **FIG. 31B** shows an example of one approach for generating molecules with methylated adenines by whole genome amplification according to embodiments of the present invention.

[0046] **FIGS. 32A** and **32B** show interpulse duration (IPD) values across sequenced A bases in template DNA of the Watson strand between unmethylated and methylated datasets according to embodiments of the present invention.

[0047] **FIG. 32C** shows a receiver operating characteristic curve for determining methylation in the Watson strand according to embodiments of the present invention.

[0048] **FIGS. 33A** and **33B** show interpulse duration (IPD) values across sequenced A bases in template DNA of the Crick strand between unmethylated and methylated datasets according to embodiments of the present invention.

[0049] **FIG. 33C** shows a receiver operating characteristic curve for determining methylation in the Crick strand according to embodiments of the present invention.

[0050] **FIG. 34** illustrates 6mA determination of the Watson strand according to embodiments of the present invention.

[0051] **FIG. 35** illustrates 6mA determination of the Crick strand according to embodiments of the present invention.

[0052] **FIGS. 36A** and **36B** show the determined probability of being methylated for sequenced A bases of the Watson strand between uA and mA datasets using a measurement window based convolutional neural network model according to embodiments of the present invention.

[0053] **FIG. 37** shows an ROC curve for the detection of 6mA using a measurement window based CNN model for sequenced A bases of the Watson strand according to embodiments of the present invention.

[0054] **FIG. 38** shows a performance comparison between IPD-metric-based 6mA detection and a measurement window based 6mA detection according to embodiments of the present invention.

[0055] **FIGS. 39A** and **39B** show the determined probability of being methylated for those sequenced A bases of the Crick strand between uA and mA datasets using measurement window based CNN model according to embodiments of the present invention.

[0056] **FIG. 40** shows the performance of 6mA detection using measurement window based CNN model on sequenced A bases of the Crick strand according to embodiments of the present invention.

[0057] **FIG. 41** shows examples of methylation states across A bases in a molecule including the Watson and Crick strands according to embodiments of the present invention.

[0058] **FIG. 42** shows an example of enhanced training by selectively using A bases in mA dataset with IPD values greater than its 10th percentile according to embodiments of the present invention.

[0059] **FIG. 43** is a graph of the percentages of unmethylated adenines in the mA dataset versus the number of subreads in each well according to embodiments of the present invention.

[0060] **FIG. 44** shows methyladenine patterns between the Watson and Crick strands of a double-stranded DNA molecule in a testing dataset according to embodiments of the present invention.

[0061] **FIG. 45** is a table showing the percentage of fully-unmethylated molecules, hemi-methylated molecules, fully-methylated molecules, and molecules with interlaced methyladenine patterns in training and testing datasets according to embodiments of the present invention.

[0062] **FIG. 46** illustrates representative examples for molecules with fully-unmethylated molecules regarding adenine sites, hemi-methylated molecules, fully-methylated molecules, and molecules with interlaced methyladenine patterns according to embodiments of the present invention.

[0063] **FIG. 47** shows an example of a long read (6,265 bp) harboring a CpG island (as shaded in yellow) according to embodiments of the present invention.

[0064] **FIG. 48** is a table showing that the 9 DNA molecules were sequenced by Pacific Biosciences SMRT sequencing and which overlapped with imprinted regions according to embodiments of the present invention.

[0065] **FIG. 49** shows an example of a genomic imprinting according to embodiments of the present invention.

[0066] **FIG. 50** shows an example for the determination of methylation patterns in an imprinted region according to embodiments of the present invention.

[0067] **FIG. 51** shows a comparison of methylation levels deduced by between the new approach and conventional bisulfite sequencing according to embodiments of the present invention.

[0068] **FIGS. 52A** and **52B** show the performance of detection of methylation of plasma DNA according to embodiments of the present invention. (A) The relationship between the predicted probability of methylation versus the ranges of methylation levels quantified by bisulfite sequencing. (B) The correlation between the methylation levels determined by Pacific Biosciences (PacBio) sequencing according to the embodiments present in this disclosure (y-axis) and methylation levels quantified by bisulfite sequencing (x-axis) in 10-Mb resolution.

[0069] **FIG. 53** shows a correlation of the genomic presentation (GR) of the Y chromosome between Pacific Biosciences SMRT sequencing and BS-seq according to embodiments of the present invention.

[0070] **FIG. 54** shows an example of CpG block based detection of methylation using CpG blocks each of which harbor a series of CpG sites according to embodiments of the present invention. 5mC: methylation; C: unmethylation.

[0071] **FIGS. 55A** and **55B** show training and testing of methylation calling for human DNA molecules using the CpG block-based approach according to embodiments of the present invention. (A) Performance in the training dataset. (B) Performance in an independent testing dataset.

[0072] **FIGS. 56A** and **56B** show copy number changes in tumor tissue according to embodiments of the present invention.

[0073] **FIGS. 57A** and **57B** show copy number changes in tumor tissue according to embodiments of the present invention.

[0074] **FIG. 58** shows a schematic illustration of plasma DNA tissue mapping from the plasma of a pregnant woman using the methylation levels deduced according to embodiments of the present invention.

[0075] **FIG. 59** shows a correlation between the placental contribution to maternal plasma DNA deduced and fetal DNA fraction deduced by Y chromosome reads according to embodiments of the present invention.

[0076] **FIG. 60** shows a table summarizing sequencing data from different human tissue DNA samples according to embodiments of the present invention.

[0077] **FIG. 61** shows an illustration of various ways of analyzing methylation patterns according to embodiments of the present invention.

[0078] **FIGS. 62A** and **62B** show a comparison of methylation densities at the whole-genome level quantified by bisulfite sequencing and single molecule, real-time sequencing according to embodiments of the present invention.

[0079] **FIGS. 63A**, **63B**, and **63C** show different correlations of overall methylation levels quantified by bisulfite sequencing and single molecule, real-time sequencing according to embodiments of the present invention.

[0080] **FIGS. 64A** and **64B** show methylation patterns at 1-Mnt resolution for a hepatocellular carcinoma (HCC) cell line and a buffy coat sample from a health control subject with methylation levels determined by bisulfite sequencing and by single molecule, real-time sequencing according to embodiments of the present invention.

[0081] **FIGS. 65A** and **65B** show scatter plots of methylation levels at a 1-Mnt resolution determined by bisulfite sequencing and single molecule, real-time sequencing according to

embodiments of the present invention for an HCC cell line (HepG2) and buffy coat sample from a healthy control subject.

[0082] **FIGS. 66A** and **66B** show scatter plots of methylation levels at a 100-knt resolution determined by bisulfite sequencing and single molecule, real-time sequencing according to embodiments of the present invention for HCC cell line (HepG2) and buffy coat sample from a healthy control subject.

[0083] **FIGS. 67A** and **67B** show methylation patterns at 1-Mnt resolution for an HCC tumor tissue and adjacent normal tissue with methylation levels determined by bisulfite sequencing and by single molecule, real-time sequencing according to embodiments of the present invention.

[0084] **FIGS. 68A** and **68B** show scatter plots of methylation levels at a 1-Mnt resolution determined by bisulfite sequencing and single molecule, real-time sequencing according to embodiments of the present invention for HCC tumor tissue and adjacent normal tissue.

[0085] **FIGS. 69A** and **69B** show scatter plots of methylation levels at a 100-knt resolution determined by bisulfite sequencing and single molecule, real-time sequencing according to embodiments of the present invention for HCC tumor tissue and adjacent normal tissue.

[0086] **FIGS. 70A** and **70B** show methylation patterns at 1-Mnt resolution for an HCC tumor tissue and adjacent normal tissue with methylation levels determined by bisulfite sequencing and by single molecule, real-time sequencing according to embodiments of the present invention.

[0087] **FIGS. 71A** and **71B** show scatter plots of methylation levels at a 1-Mnt resolution determined by bisulfite sequencing and single molecule, real-time sequencing according to embodiments of the present invention for HCC tumor tissue and adjacent normal tissue.

[0088] **FIGS. 72A** and **72B** show scatter plots of methylation levels at a 100-knt resolution determined by bisulfite sequencing and single molecule, real-time sequencing according to embodiments of the present invention for HCC tumor tissue and adjacent normal tissue.

[0089] **FIG. 73** shows an example of the aberrant pattern of methylation nearby the tumor suppressor gene *CDKN2A* according to embodiments of the present invention.

[0090] **FIGS. 74A** and **74B** show differential methylation regions detected by single molecule, real-time sequencing according to embodiments of the present invention.

[0091] **FIG. 75** shows methylation patterns of hepatitis B virus DNA between HCC tissues and adjacent non-tumoral tissues using single molecule, real-time sequencing according to embodiments of the present invention.

[0092] **FIG. 76A** shows methylation levels of hepatitis B virus DNA in liver tissues from patients with cirrhosis but without HCC using bisulfite sequencing according to embodiments of the present invention.

[0093] **FIG. 76B** shows methylation levels of hepatitis B virus DNA in HCC tissues using bisulfite sequencing according to embodiments of the present invention.

[0094] **FIG. 77** illustrates methylation haplotype analysis according to embodiments of the present invention.

[0095] **FIG. 78** shows the size distribution of the sequenced molecules determined from consensus sequences according to embodiments of the present invention.

[0096] **FIGS. 79A**, **79B**, **79C**, and **79D** show examples of allelic methylation patterns in the imprinted regions according to embodiments of the present invention.

[0097] **FIGS. 80A**, **80B**, **80C**, and **80D** show examples of allelic methylation patterns in non-imprinted regions according to embodiments of the present invention.

[0098] **FIG. 81** shows a table of methylation levels of allele-specific fragments according to embodiments of the present invention.

[0099] **FIG. 82** shows an example for determining the placental origin of plasma DNA in pregnancy using methylation profiles according to embodiments of the present invention.

[0100] **FIG. 83** illustrates fetal-specific DNA methylation analysis according to embodiments of the present invention.

[0101] **FIGS. 84A**, **84B**, and **84C** show the performance of different measurement window sizes across different reagent kits for SMRT-seq according to embodiments of the present invention.

[0102] **FIGS. 85A**, **85B**, and **85C** show the performance of different measurement window sizes across different reagent kits for SMRT-seq according to embodiments of the present invention.

[0103] **FIGS. 86A**, **86B**, and **86C** show the correlation of overall methylation levels quantified by bisulfite sequencing and SMRT-seq (Sequel II Sequencing Kit 2.0) according to embodiments of the present invention.

[0104] **FIGS. 87A** and **87B** show a comparison of overall methylation level between various tumor tissues and paired adjacent non-tumoral tissues according to embodiments of the present invention.

[0105] **FIG. 88** shows determining the methylation status using a sequence context determined from a circular consensus sequence (CCS) according to embodiments of the present invention.

[0106] **FIG. 89** shows an ROC curve for the detection of methylated CpG sites using a sequence context determined from CCS according to embodiments of the present invention.

[0107] **FIG. 90** shows an ROC curve for the detection of methylated CpG sites without CCS information and without prior alignment to a reference genome according to embodiments of the present invention.

[0108] **FIG. 91** shows an example of preparing molecules for single molecule, real-time sequencing according to embodiments of the present invention.

[0109] **FIG. 92** shows an illustration of the CRISPR/Cas9 system according to embodiments of the present invention.

[0110] **FIG. 93** shows an example of a Cas9 complex for introducing two cuts spanning an end-blocked molecule of interest according to embodiments of the present invention.

[0111] **FIG. 94** shows methylation distribution of Alu regions determined by bisulfite sequencing and single molecule, real-time sequencing according to embodiments of the present invention.

[0112] **FIG. 95** shows the distribution of methylation levels of Alu regions determined by model using results from single molecule, real-time sequencing according to embodiments of the present invention.

[0113] **FIG. 96** shows a table of tissues and the methylation levels of Alu regions in the tissues according to embodiments of the present invention.

[0114] **FIG. 97** shows clustering analysis for different cancer types using methylation signals related to Alu repeats according to embodiments of the present invention.

[0115] **FIG. 98A** and **98B** show the effect of read depth on the overall methylation level quantification in the testing datasets which were involved with the whole genome amplification and M.SsssI treatment according to embodiments of the present invention.

[0116] **FIG. 99** shows a comparison between overall methylation levels determined by SMRT-seq (Sequel II Sequencing Kit 2.0) and BS-seq with the use of different subread depth cutoffs according to embodiments of the present invention.

[0117] **FIG. 100** is a table showing the effect of the subread depth on the correlation of methylation levels between two measurements by SMRT-seq (Sequel II Sequencing Kit 2.0) and BS-seq according to embodiments of the present invention.

[0118] **FIG. 101** shows the subread depth distribution with respect to fragment sizes in data generated by Sequel II Sequencing Kit 2.0 according to embodiments of the present invention.

[0119] **FIG. 102** shows a method of detecting a modification of a nucleotide in a nucleic acid molecule according to embodiments of the present invention.

[0120] **FIG. 103** shows a method for detecting a modification of a nucleotide in a nucleic acid molecule according to embodiments of the present invention.

[0121] **FIG. 104** illustrates relative haplotype-based methylation imbalance analysis according to embodiments of the present invention.

[0122] **FIG. 105A** and **105B** are a table of the haplotype blocks showing differential methylation levels between Hap I and Hap II in the tumor DNA compared with the adjacent non-tumoral tissue DNA for the case TBR3033 according to embodiments of the present invention.

[0123] **FIG. 106** is a table of the haplotype blocks showing differential methylation levels between Hap I and Hap II in the tumor DNA compared with the adjacent normal tissue DNA for the case TBR3032 according to embodiments of the present invention.

[0124] **FIG. 107A** is a table summarizing the number of haplotype blocks showing methylation imbalance between two haplotypes between tumor and adjacent nontumoral tissues on the basis of data generated by Sequel II Sequencing Kit 2.0 according to embodiments of the present invention.

[0125] **FIG. 107B** is a table summarizing the number of haplotype blocks showing methylation imbalance between two haplotypes in tumor tissues for different tumor stages on the basis of data generated by Sequel II Sequencing Kit 2.0 according to embodiments of the present invention.

[0126] **FIG. 108** illustrates relative haplotype-based methylation imbalance analysis according to embodiments of the present invention.

[0127] **FIG. 109** shows a method of classifying a disorder in an organism having a first haplotype and a second haplotype according to embodiments of the present invention.

[0128] **FIG. 110** illustrates creating human-mouse hybrid fragments for which the human part is methylated while the mouse part is unmethylated according to embodiments of the present invention.

[0129] **FIG. 111** illustrates creating human-mouse hybrid fragments for which the human part is unmethylated while the mouse part is methylated according to embodiments of the present invention.

[0130] **FIG. 112** shows the length distribution of DNA molecules in a DNA mixture (sample MIX01) after ligation according to embodiments of the present invention.

[0131] **FIG. 113** illustrates a junction region by which a first DNA (A) and a second DNA (B) are joined together according to embodiments of the present invention.

[0132] **FIG. 114** illustrates methylation analysis for the DNA mixture according to embodiments of the present invention.

[0133] **FIG. 115** shows a boxplot of the probabilities of being methylated for CpG sites in the sample MIX01 according to embodiments of the present invention.

[0134] **FIG. 116** shows length distribution of DNA molecules in the DNA mixture after cross-ligation of sample MIX02 according to embodiments of the present invention.

[0135] **FIG. 117** shows a boxplot of the probabilities of being methylated for CpG sites in the sample MIX02 according to embodiments of the present invention.

[0136] **FIG. 118** is a table comparing methylation determined by bisulfite sequencing and Pacific Biosciences sequencing for MIX01 according to embodiments of the present invention.

[0137] **FIG. 119** is a table comparing methylation determined by bisulfite sequencing and Pacific Biosciences sequencing for MIX02 according to embodiments of the present invention.

[0138] **FIGS. 120A** and **120B** show methylation levels in 5-Mb bins for human-only and mouse-only DNA for MIX01 and MIX02 according to embodiments of the present invention.

[0139] **FIGS. 121A** and **121B** show methylation levels in 5-Mb bins for the human part and the mouse part of human-mouse hybrid DNA fragments for MIX01 and MIX02 according to embodiments of the present invention.

[0140] **FIGS. 122A** and **122B** are representative graphs showing methylation states in a single human-mouse hybrid molecule according to embodiments of the present invention.

[0141] **FIG. 123** shows a method of detecting chimeric molecules in a biological sample according to embodiments of the present invention.

[0142] **FIG. 124** illustrates a measurement system according to embodiments of the present invention.

[0143] **FIG. 125** shows a block diagram of an example computer system usable with systems and methods according to embodiments of the present invention.

[0144] **FIG. 126** shows *Msp*I-based targeted single molecule, real-time sequencing with the use of DNA end repair and A-tailing according to embodiments of the present invention.

[0145] **FIGS. 127A** and **127B** show the size distribution of *Msp*I-digested fragments according to embodiments of the present invention.

[0146] **FIG. 128** shows a table with the number of DNA molecules for certain selected size ranges according to embodiments of the present invention.

[0147] **FIG. 129** is a graph of the percentage coverage of CpG sites within CpG islands versus size of DNA fragments after restriction enzyme digestion according to embodiments of the present invention.

[0148] **FIG. 130** shows *Msp*I-based targeted single molecule, real time sequencing without the use of DNA end repair and A-tailing according to embodiments of the present invention.

[0149] **FIG. 131** shows *Msp*I-based targeted single molecule, real time sequencing with a reduced probability of adapter self-ligation according to embodiments of the present invention.

[0150] **FIG. 132** is a graph of the overall methylation levels between placenta and buffy DNA samples determined by *Msp*I-based targeted single molecule, real-time sequencing according to the embodiments of the present invention.

[0151] **FIG. 133** shows a clustering analysis of placental and buffy coat samples using their DNA methylation profiles determined by *Msp*I-based targeted single molecule, real-time sequencing according to the embodiments of the present invention.

TERMS

[0152] A "*tissue*" corresponds to a group of cells that group together as a functional unit. More than one type of cells can be found in a single tissue. Different types of tissue may consist of

different types of cells (e.g., hepatocytes, alveolar cells or blood cells), but also may correspond to tissue from different organisms (mother vs. fetus; tissues in a subject who has received transplantation; tissues of an organism that are infected by a microorganism or a virus) or to healthy cells vs. tumor cells. "Reference tissues" can correspond to tissues used to determine tissue-specific methylation levels. Multiple samples of a same tissue type from different individuals may be used to determine a tissue-specific methylation level for that tissue type.

[0153] A "*biological sample*" refers to any sample that is taken from a human subject. The biological sample can be a tissue biopsy, a fine needle aspirate, or blood cells. The sample can also be for example, plasma or serum or urine from a pregnant woman. Stool samples can also be used. In various embodiments, the majority of DNA in a biological sample from a pregnant woman that has been enriched for cell-free DNA (e.g., a plasma sample obtained via a centrifugation protocol) can be cell-free, e.g., greater than 50%, 60%, 70%, 80%, 90%, 95%, or 99% of the DNA can be cell-free. The centrifugation protocol can include, for example, 3,000 g x 10 minutes, obtaining the fluid part, and re-centrifuging at for example, 30,000 g for another 10 minutes to remove residual cells. In certain embodiments, following the 3,000 g centrifugation step, one can follow up with filtration of the fluid part (e.g. using a filter of pore size of 5 μm, or smaller, in diameter).

[0154] A "*sequence read*" refers to a string of nucleotides sequenced from any part or all of a nucleic acid molecule. For example, a sequence read may be a short string of nucleotides (e.g., 20-150) sequenced from a nucleic acid fragment, a short string of nucleotides at one or both ends of a nucleic acid fragment, or the sequencing of the entire nucleic acid fragment that exists in the biological sample. A sequence read may be obtained in a variety of ways, e.g., using sequencing techniques or using probes, e.g., in hybridization arrays or capture probes, or amplification techniques, such as the polymerase chain reaction (PCR) or linear amplification using a single primer or isothermal amplification.

[0155] A "*subread*" is a sequence generated from all bases in one strand of a circularized DNA template that has been copied in one contiguous strand by a DNA polymerase. For example, a subread can correspond to one strand of circularized DNA template DNA. In such an example,

after circularization, one double-stranded DNA molecule would have two subreads: one for each sequencing pass. In some embodiments, the sequence generated may include a subset of all the bases in one strand, e.g., because of the existence of sequencing errors.

[0156] A "*site*" (also called a "*genomic site*") corresponds to a single site, which may be a single base position or a group of correlated base positions, e.g., a CpG site or larger group of correlated base positions. A "locus" may correspond to a region that includes multiple sites. A locus can include just one site, which would make the locus equivalent to a site in that context.

[0157] A "*methylation status*" refers to the state of methylation at a given site. For example, a site may be either methylated, unmethylated, or in some cases, undetermined.

[0158] The "*methylation index*" for each genomic site (e.g., a CpG site) can refer to the proportion of DNA fragments (e.g., as determined from sequence reads or probes) showing methylation at the site over the total number of reads covering that site. A "read" can correspond to information (e.g., methylation status at a site) obtained from a DNA fragment. A read can be obtained using reagents (e.g. primers or probes) that preferentially hybridize to DNA fragments of a particular methylation status at one or more sites. Typically, such reagents are applied after treatment with a process that differentially modifies or differentially recognizes DNA molecules depending on their methylation status, e.g. bisulfite conversion, or methylation-sensitive restriction enzyme, or methylation binding proteins, or anti-methylcytosine antibodies, or single molecule sequencing techniques (e.g. single molecule, real-time sequencing and nanopore sequencing (e.g. from Oxford Nanopore Technologies)) that recognize methylcytosines and hydroxymethylcytosines.

[0159] The "*methylation density*" of a region can refer to the number of reads at sites within the region showing methylation divided by the total number of reads covering the sites in the region. The sites may have specific characteristics, e.g., being CpG sites. Thus, the "CpG methylation density" of a region can refer to the number of reads showing CpG methylation divided by the total number of reads covering CpG sites in the region (e.g., a particular CpG site, CpG sites within a CpG island, or a larger region). For example, the methylation density for each 100-kb bin in the human genome can be determined from the total number of cytosines not

converted after bisulfite treatment (which corresponds to methylated cytosine) at CpG sites as a proportion of all CpG sites covered by sequence reads mapped to the 100-kb region. This analysis can also be performed for other bin sizes, e.g. 500 bp, 5 kb, 10 kb, 50-kb or 1-Mb, etc. A region could be the entire genome or a chromosome or part of a chromosome (e.g. a chromosomal arm). The methylation index of a CpG site is the same as the methylation density for a region when the region only includes that CpG site. The "proportion of methylated cytosines" can refer the number of cytosine sites, "C's", that are shown to be methylated (for example unconverted after bisulfite conversion) over the total number of analyzed cytosine residues, i.e. including cytosines outside of the CpG context, in the region. The methylation index, methylation density, count of molecules methylated at one or more sites, and proportion of molecules methylated (e.g., cytosines) at one or more sites are examples of "*methylation levels*." Apart from bisulfite conversion, other processes known to those skilled in the art can be used to interrogate the methylation status of DNA molecules, including, but not limited to enzymes sensitive to the methylation status (e.g. methylation-sensitive restriction enzymes), methylation binding proteins, single molecule sequencing using a platform sensitive to the methylation status (e.g. nanopore sequencing (Schreiber et al. Proc Natl Acad Sci 2013; 110: 18910-18915) and by single molecule, real-time sequencing (e.g. that from Pacific Biosciences) (Flusberg et al. Nat Methods 2010; 7: 461-465)).

[0160] A "*methylome*" provides a measure of an amount of DNA methylation at a plurality of sites or loci in a genome. The methylome may correspond to all of the genome, a substantial part of the genome, or relatively small portion(s) of the genome.

[0161] A "*pregnant plasma methylome*" is the methylome determined from the plasma or serum of a pregnant animal (e.g., a human). The pregnant plasma methylome is an example of a cell-free methylome since plasma and serum include cell-free DNA. The pregnant plasma methylome is also an example of a mixed methylome since it is a mixture of DNA from different organs or tissues or cells within a body. In one embodiment, such cells are the hematopoietic cells, including, but not limited to cells of the erythroid (i.e. red cell) lineage, the myeloid lineage (e.g., neutrophils and their precursors), and the megakaryocytic lineage. In pregnancy, the plasma methylome may contain methylomic information from the fetus and the mother. The

"cellular methylome" corresponds to the methylome determined from cells (e.g., blood cells) of the patient. The methylome of the blood cells is called the blood cell methylome (or blood methylome).

[0162] A "*methylation profile*" includes information related to DNA or RNA methylation for multiple sites or regions. Information related to DNA methylation can include, but not limited to, a methylation index of a CpG site, a methylation density (MD for short) of CpG sites in a region, a distribution of CpG sites over a contiguous region, a pattern or level of methylation for each individual CpG site within a region that contains more than one CpG site, and non-CpG methylation. In one embodiment, the methylation profile can include the pattern of methylation or non-methylation of more than one type of base (e.g. cytosine or adenine). A methylation profile of a substantial part of the genome can be considered equivalent to the methylome. "DNA methylation" in mammalian genomes typically refers to the addition of a methyl group to the 5' carbon of cytosine residues (i.e. 5-methylcytosines) among CpG dinucleotides. DNA methylation may occur in cytosines in other contexts, for example CHG and CHH, where H is adenine, cytosine or thymine. Cytosine methylation may also be in the form of 5-hydroxymethylcytosine. Non-cytosine methylation, such as N^6-methyladenine, has also been reported.

[0163] A "*methylation pattern*" refers to the order of methylated and non-methylated bases. For example, the methylation pattern can be the order of methylated bases on a single DNA strand, a single double-stranded DNA molecule, or another type of nucleic acid molecule. As an example, three consecutive CpG sites may have any of the following methylation patterns: UUU, MMM, UMM, UMU, UUM, MUM, MUU, or MMU, where "U" indicates an unmethylated site and "M" indicates a methylated site. When one extends this concept to base modifications that include, but not restricted to methylation, one would use the term "*modification pattern*," which refers to the order of modified and non-modified bases. For example, the modification pattern can be the order of modified bases on a single DNA strand, a single double-stranded DNA molecule, or another type of nucleic acid molecule. As an example, three consecutive potentially modifiable sites may have any of the following modification patterns: UUU, MMM, UMM, UMU, UUM, MUM, MUU, or MMU, where "U" indicates an unmodified site and "M" indicates

a modified site. One example of base modification that is not based on methylation is oxidation changes, such as in 8-oxo-guanine.

[0164] The terms "*hypermethylated*" and "*hypomethylated*" may refer to the methylation density of a single DNA molecule as measured by its single molecule methylation level, e.g., the number of methylated bases or nucleotides within the molecule divided by the total number of methylatable bases or nucleotides within that molecule. A hypermethylated molecule is one in which the single molecule methylation level is at or above a threshold, which may be defined from application to application. The threshold may be 5%, 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90%, or 95%. A hypomethylated molecule is one in which the single molecule methylation level is at or below a threshold, which may be defined from application to application, and which may change from application to application. The threshold may be 5%, 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90%, or 95%.

[0165] The terms "hypermethylated" and "hypomethylated" may also refer to the methylation level of a population of DNA molecules as measured by the multiple molecule methylation levels of these molecules. A hypermethylated population of molecules is one in which the multiple molecule methylation level is at or above a threshold which may be defined from application to application, and which may change from application to application. The threshold may be 5%, 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90%, or 95%. A hypomethylated population of molecules is one in which the multiple molecule methylation level is at or below a threshold which may be defined from application to application. The threshold may be 5%, 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90%, and 95%. In one embodiment, the population of molecules may be aligned to one or more selected genomic regions. In one embodiment, the selected genomic region(s) may be related to a disease such as cancer, a genetic disorder, an imprinting disorder, a metabolic disorder, or a neurological disorder. The selected genomic region(s) can have a length of 50 nucleotides (nt), 100 nt, 200 nt, 300 nt, 500 nt, 1000 nt, 2 knt, 5 knt, 10 knt, 20 knt, 30 knt, 40 knt, 50 knt, 60 knt, 70 knt, 80 knt, 90 knt, 100 knt, 200 knt, 300 knt, 400 knt, 500 knt, or 1 Mnt.

[0166] The term "*sequencing depth*" refers to the number of times a locus is covered by a sequence read aligned to the locus. The locus could be as small as a nucleotide, or as large as a chromosome arm, or as large as the entire genome. Sequencing depth can be expressed as 50x, 100x, etc., where "x" refers to the number of times a locus is covered with a sequence read. Sequencing depth can also be applied to multiple loci, or the whole genome, in which case x can refer to the mean number of times the loci or the haploid genome, or the whole genome, respectively, is sequenced. Ultra-deep sequencing can refer to at least 100x in sequencing depth.

[0167] The term "*classification*" as used herein refers to any number(s) or other characters(s) that are associated with a particular property of a sample. For example, a "+" symbol (or the word "positive") could signify that a sample is classified as having deletions or amplifications. The classification can be binary (e.g., positive or negative) or have more levels of classification (e.g., a scale from 1 to 10 or 0 to 1).

[0168] The terms "*cutoff*" and "*threshold*" refer to predetermined numbers used in an operation. For example, a cutoff size can refer to a size above which fragments are excluded. A threshold value may be a value above or below which a particular classification applies. Either of these terms can be used in either of these contexts. A cutoff or threshold may be "a reference value" or derived from a reference value that is representative of a particular classification or discriminates between two or more classifications. Such a reference value can be determined in various ways, as will be appreciated by the skilled person. For example, metrics can be determined for two different cohorts of subjects with different known classifications, and a reference value can be selected as representative of one classification (e.g., a mean) or a value that is between two clusters of the metrics (e.g., chosen to obtain a desired sensitivity and specificity). As another example, a reference value can be determined based on statistical analyses or simulations of samples.

[0169] The term "*level of cancer*" can refer to whether cancer exists (i.e., presence or absence), a stage of a cancer, a size of tumor, whether there is metastasis, the total tumor burden of the body, the cancer's response to treatment, and/or other measure of a severity of a cancer (e.g. recurrence of cancer). The level of cancer may be a number or other indicia, such as symbols,

alphabet letters, and colors. The level may be zero. The level of cancer may also include premalignant or precancerous conditions (states). The level of cancer can be used in various ways. For example, screening can check if cancer is present in someone who is not previously known to have cancer. Assessment can investigate someone who has been diagnosed with cancer to monitor the progress of cancer over time, study the effectiveness of therapies or to determine the prognosis. In one embodiment, the prognosis can be expressed as the chance of a patient dying of cancer, or the chance of the cancer progressing after a specific duration or time, or the chance or extent of cancer metastasizing. Detection can mean 'screening' or can mean checking if someone, with suggestive features of cancer (e.g. symptoms or other positive tests), has cancer.

[0170] A "*level of pathology*" (or level of disorder) can refer to the amount, degree, or severity of pathology associated with an organism, where the level can be as described above for cancer. Another example of pathology is a rejection of a transplanted organ. Other example pathologies can include gene imprinting disorders, autoimmune attack (e.g., lupus nephritis damaging the kidney or multiple sclerosis), inflammatory diseases (e.g., hepatitis), fibrotic processes (e.g. cirrhosis), fatty infiltration (e.g. fatty liver diseases), degenerative processes (e.g. Alzheimer's disease), and ischemic tissue damage (e.g., myocardial infarction or stroke). A heathy state of a subject can be considered a classification of no pathology.

[0171] A "*pregnancy-associated disorder*" include any disorder characterized by abnormal relative expression levels of genes in maternal and/or fetal tissue. These disorders include, but are not limited to, preeclampsia, intrauterine growth restriction, invasive placentation, pre-term birth, hemolytic disease of the newborn, placental insufficiency, hydrops fetalis, fetal malformation, HELLP syndrome, systemic lupus erythematosus, and other immunological diseases of the mother.

[0172] The abbreviation "*bp*" refers to base pairs. In some instances, "bp" may be used to denote a length of a DNA fragment, even though the DNA fragment may be single stranded and does not include a base pair. In the context of single-stranded DNA, "bp" may be interpreted as providing the length in nucleotides.

[0173] The abbreviation "nt" refers to nucleotides. In some instances, "nt" may be used to denote a length of a single-standed DNA in a base unit. Also, "nt" may be used to denote the relative positions such as upstream or downstream of the locus being analyzied. In some contexts concerning technological conceptualization, data presentation, processing and analysis, "nt" and "bp" may be used interchangeably.

[0174] The term "*sequence context*" can refer to the base compositions (A, C, G, or T) and the base orders in a stretch of DNA. Such a stretch of DNA could be surrounding a base that is subjected to or the target of base modification analysis. For example, the seqeunce context can refer to bases upstream and/or downstream of a base that is subjected to base modification analysis.

[0175] The term "*kinetic features*" can refer to features derived from sequencing, including from single molecule, real-time sequencing. Such features can be used for base modification analysis. Example kinetic features include upstream and downstream sequence context, strand information, interpulse duration, pulse widths, and pulse strength. In single molecule, real-time sequencing, one is continuously monitoring the effects of activities of a polymerase on a DNA template. Hence, measurements generated from such a sequencing can be regarded as kinetic features, e.g., nucleotide sequences.

[0176] The term "*machine learning models*" may include models based on using sample data (e.g., training data) to make predictions on test data, and thus may include supervised learning. Machine learning models often are developed using a computer or a processor. Machine learning models may include statistical models.

[0177] The term "*data analysis framework*" may include algorithms and/or models that can take data as an input and then output a predicted result. Examples of "data analysis frameworks" include statistical models, mathematical models, machine learning models, other artificial intelligence models, and combinations thereof.

[0178] The term "*real-time sequencing*" may refer to a technique that involves data collection or monitoring during progress of a reaction involved in sequencing. For example, real-time

sequencing may involve optical monitoring or filming the DNA polymerase incorporatingincorporate a new base.

[0179] The term "*about*" or "*approximately*" can mean within an acceptable error range for the particular value as determined by one of ordinary skill in the art, which will depend in part on how the value is measured or determined, i.e., the limitations of the measurement system. For example, "about" can mean within 1 or more than 1 standard deviation, per the practice in the art. Alternatively, "about" can mean a range of up to 20%, up to 10%, up to 5%, or up to 1% of a given value. Alternatively, particularly with respect to biological systems or processes, the term "about" or "approximately" can mean within an order of magnitude, within 5-fold, and more preferably within 2-fold, of a value. Where particular values are described in the application and claims, unless otherwise stated the term "about" meaning within an acceptable error range for the particular value should be assumed. The term "about" can have the meaning as commonly understood by one of ordinary skill in the art. The term "about" can refer to ±10%. The term "about" can refer to ±5%.

DETAILED DESCRIPTION

[0180] Achieving bisulfite-free determination of a base modification, include a methylated base, is the subject of different research efforts but none have been shown to be commercially viable. Recently, a bisulfite-free method for detecting 5mC and 5hmC has been published (Y. Liu *et al.*, 2019) using a mild condition for 5mC and 5hmC base conversion. This method involves multiple steps of enzymatic and chemical reactions including ten-eleven translocation (TET) oxidation, pyridine borane reduction, and PCR. The efficiency for each step of conversion reaction as well as PCR bias would adversely affect the ultimate accuracy in 5mC analysis. For example, the 5mC conversion rate has been reported to be around 96%, with a false-negative rate of around 3%. Such performance would potentially limit one's ability to detect certain subtle changes of methylation in a genome. On the other hand, the enzymatic conversion would not be able to perform equally well across the genome. For example, the conversion rate of 5hmC was 8.2% lower than that for 5mC, and the conversion rate for non-CpG was 11.4% lower than that for CpG contexts (Y. Liu *et al.*, 2019). Thus, the ideal situation is the development of approaches

for measuring base modifications of a native DNA molecule without any prior conversion (chemical or enzymatic, or combinations thereof) step and even without an amplification step.

[0181] There were a number of proof-of-concept studies (Q. Liu *et al.*, 2019; Ni *et al.*, 2019) in which the electric signals produced by a long-read nanopore sequencing approach (e.g., using the system developed by Oxford Nanopore Technologies) enabled one to detect methylation states with the use of a deep learning method. In addition to Oxford Nanopore, there are other single molecule sequencing approaches that allow long reads. One example is the single molecule, real-time sequencing. One example of a single molecule, real-time sequencing is that commercialized the Pacific Biosciences SMRT system. As the principle of single molecule, real-time sequencing (e.g., the Pacific Biosciences SMRT system) is different from that of a non-optical based nanopore system (e.g. by Oxford Nanopore Technologies), approaches for base modification detection developed for such non-optical based nanopore system cannot be used for single molecule, real-time sequencing. For example, a non-optical nanopore system is not designed for capturing the patterns of fluorescent signals produced by immobilized DNA polymerase based DNA synthesis (employed by single molecule, real-time sequencing such as by the Pacific Biosciences SMRT system). As a further example, in the Oxford Nanopore sequencing platform, each measured electric event is associated with a k-mer (e.g., 5-mer) (Q. Liu *et al.*, 2019). However, in the Pacific Biosciences SMRT sequencing platform, each fluorescent event is generally associated with a single incorporated base. Furthermore, a single DNA molecule would be sequenced multiple times in Pacific Biosciences SMRT sequencing including Watson and Crick strands. Conversely, for the Oxford Nanopore long-read sequencing approach, sequence readout is performed once for each of the Watson and Crick strands.

[0182] It has been reported that the polymerase kinetics would be affected by methylation states in the sequences of *E. coli* (Flusberg *et al.*, 2010). Previous studies showed that when compared with the detection of 6mA, 4mC, 5hmC, and 8-oxo-guanine, it is much more challenging to use the polymerase kinetics of single molecule, real-time sequencing for deducing the methylation states (5mC versus C) of a particular CpG in a single molecule. The reason is that the methyl group is small and oriented towards the major groove and is not involved in base pairing, leading to very subtle interruption in the kinetics caused by 5mC (Clark *et al.*, 2013).

Hence, there is a paucity of approaches for determining the methylation states of cytosines at the single-molecule level.

[0183] Suzuki et al developed an algorithm (Suzuki *et al.*, 2016) attempting to combine the interpulse duration (IPD) ratios for neighboring CpG sites to increase the confidence in identifying the methylation states of those sites. However, this algorithm only allowed one to predict a genomic region of being completely methylated or completely unmethylated, but lacked the ability to determine intermediate methylation patterns.

[0184] Regarding single molecule, real-time sequencing, current approaches only used one or two parameters independently, achieving a very limited accuracy in detecting 5mC because of the measurement difference between 5-methylcytosine and cytosine. For example, Flusberg *et al.* demonstrated that IPD was altered in base modifications including N6-methyladenosine, 5-methylcytosine, and 5-hydroxymethylcytosine. However, pulse width (PW) of the sequencing kinetics was not found to have a significant effect. Hence, in the method they used for predicting base modification, using the detection of N6-methyladenosine as an example, only IPD but not PW was used.

[0185] In follow-up publications by the same group (Clark *et al.*, 2012; Clark *et al.* 2013), IPD but not PW was incorporated in the algorithms for the detection of 5-methylcytosine. In Clark *et al.* 2012, the detection rate of 5-methylcytosine without converting it to 5-methylcytosine only ranged from 1.9% to 4.3%. Furthermore, in Clark *et al.* 2013, the authors had further reaffirmed the subtlety of the kinetic signature of 5-methylcytosine. To overcome the low sensitivity of detecting 5-methylcytosine, Clark *et al.* further developed a method which converted 5-methylcytosine to 5-carboxylmethylcytosine using Ten-eleven translocation (Tet) proteins so as to improve the sensitivity of 5-methylcytosine (Clark *et al.* 2013) because the alteration of IPD caused by 5-carboxylcytosine was much more than by 5-methylcytosine.

[0186] In a more recent report by Blow *et al.,* the IPD ratio-based method previously described by Flusberg *et al* was used to detect the base modifications in 217 bacterial and 13 archaeal species with 130-fold read coverage per organism (Blow *et al.*, 2016). Among all the base modifications they identified, only 5% involved 5-methylcytosine. They attributed this low

detection rate of 5-methylcytosine to the low sensitivity of single-molecule real-time sequencing for detecting 5-methylcytosine. In most bacteria, a set of sequence motifs were targeted by DNA methyltransferases (MTases) for methylation (e.g. 5'-GmATC-3' by Dam or 5'-CmCWGG-3' by Dcm in *E. coli*) at nearly *all* of these motifs in the genome, with only a small fraction of these motif sites remaining non-methylated (Beaulaurier *et al.* 2019). Furthermore, the use of the IPD-based method to classify the methylation status of the second C in the 5'-CCWGG-3' motif with, or without treatment, with Tet proteins yielded detection rates of 5-methylcytosine of 95.2% and 1.9%, respectively (Clark *et al.* 2013). Taken as a whole, the IPD method without prior base conversion (e.g., using Tet proteins) missed the majority of 5-methylcytosine.

[0187] In the studies mentioned above (Clark *et al.*, 2012; Clark *et al.*, 2013; Blow *et al.*, 2016), IPD-based algorithms were used without consideration of the sequence context at where the candidate base modification was located. Other groups have attempted to take into account the sequence context of a nucleotide for the detection of base modification. For example, Feng *et al.* used a hierarchical model to analyze IPDs for the detection of 4-methylcytosine and 6-methyladenosine in a respective sequence context (Feng *et al.* 2013). However, in their method, they only considered the IPD at the base of interest and the sequence context adjacent to that base, but did not use the IPD information of all neighboring bases adjacent to the base of interest. In addition, PW was not considered in the algorithm, and they did not present any data on the detection of 5-methylcytosine.

[0188] In another study, Schadt *et al.* developed a statistical method, called conditional random field, to analyze the IPD information of the base of interest and the neighboring bases to determine if the base of interest was a 5-methylcytosine (Schadt *et al.*, 2012). In this work, they also considered the IPD interaction between these bases by inputting them into an equation. However, they did not input the nucleotide sequence, namely A, T, G, or C, in their equation. When they applied the method to determine the methylation status of the M.Sau3AI plasmid, the area under ROC curve was close to 0.5 even at an 800-fold sequence coverage of the plasmid sequence. Moreover, in their method, they had not taken into account PW in their analysis.

[0189] In yet another study by Beckman *et al.*, they compared the IPD of all sequences that shared the same 4-nt or 6-nt motif in the genome between a target bacterial genome and a completely unmethylated genome, e.g., obtained through whole genome amplification (Beckman *et al.* 2014). The purpose of such analysis was only to identify motifs that would be more frequently affected by base modifications. In the study, they only considered the IPD of a potentially modified base but not the IPD of the neighboring base or PW. Their method was not informative about the methylation status of individual nucleotide.

[0190] In summary, these previous attempts of utilizing IPD only or with combination of sequence information in the neighboring nucleotides for grouping data were not able to determine the base modification of 5-methylcytosine with meaningful or practical accuracy. In a recent review by Gouil *et al.*, the authors concluded that because of the low signal-to-noise ratio, the detection of 5-methylcytosine in a single molecule using single-molecule real-time sequencing is inaccurate (Gouil *et al.*, 2019). In these previous studies, it remains unknown as to whether it may be feasible to use the kinetic features for genomewide methylomic analysis, especially for complex genomes such as human genomes, cancer genomes, or fetal genomes.

[0191] In contrast to previous studies, some embodiments of methods described in this disclosure are based on measuring and utilizing IPD, PW, and sequence context for every base within the measurement window. We reasoned that if we can use a combination of multiple metrics, for example, concurrently making use of features including upstream and downstream sequence context, strand information, IPD, pulse widths as well as pulse strength, we might be able to achieve the accurate measurement of base modifications (e.g. mC detection) at single-base resolution. Sequence context refers to the base compositions (A, C, G, or T) and the base orders in a stretch of DNA. Such a stretch of DNA could be surrounding a base that is subjected to or the target of base modification analysis. In one embodiment, the stretch of DNA could be proximal to a base that is subjected to base modification analysis. In another embodiment, the stretch of DNA could be far away from a base that is subjected to base modification analysis. The stretch of DNA could be upstream and/or downstream of a base that is subjected to base modification analysis.

[0192] In one embodiment, the features of upstream and downstream sequence context, strand information, IPD, pulse widths as well as pulse strength, which are used for base modification analysis, are referred to as kinetic features.

[0193] The embodiments present in this disclosure can be used for DNA obtained from, but not limited to, cell lines, samples from an organism (e.g. solid organs, solid tissues, a sample obtained via endoscopy, blood, or plasma or serum or urine from a pregnant woman, chorionic villus biopsy, etc.), samples obtained from the environment (e.g. bacteria, cellular contaminants), food (e.g., meat). In some embodiments, the methods present in this disclosure can also be applied following a step in which a fraction of the genome is first enriched, e.g. using hybridization probes (Albert *et al.*, 2007; Okou *et al.*, 2007; Lee *et al.*, 2011), or approaches based on physical separation (e.g. based on sizes, etc) or following restriction enzyme digestion (e.g. MspI), or Cas9-based enrichment (Watson *et al.*, 2019). While the invention does not require enzymatic or chemical conversion to work, in certain embodiments, such a conversion step can be included to further enhance the performance of the invention.

[0194] Embodiments of the present disclosure allow for improved accuracy or practicality or convenience in detecting base modifications or measuring modification levels. The modification may be detected directly. Embodiments may avoid enzymatic or chemical conversion, which may not preserve all modification information for detection. Additionally, certain enzymatic or chemical conversions may not be compatible with certain types of modifications. Embodiments of the present disclosure may also avoid amplification by PCR, which may not transfer base-modification information to the PCR products. Additionally, both strands of DNA may be sequenced together, thereby enabling the pairing of the sequence from one strand with its complementary sequence to the other strand. By contrast, PCR amplification splits the two strands of double-stranded DNA, so such pairing of sequences is difficult.

[0195] Methylation profiles, determined with or without enzymatic or chemical conversion, can be used for analyzing biological samples. In one embodiment, the methylation profiles can be used to detect the origin of cellular DNA (e.g., maternal or fetal, tissue, viral, or tumor). Detection of aberrant methylation profiles in tissues aid the identification of developmental

disorders in individuals and the identification and prognostication of tumors or malignancies. Imbalances in methylation levels between hapltoypes can be used to detect disorders, including cancer. Methylation patterns in a single molecule can identify chimeric (e.g., between a virus and human) and hybrid DNA, (e.g., between two genes normally unfused in a natural genome); or between two species (e.g., through genetic or genomic manipulation).

[0196] Methylation analysis may be improved by enhanced training, which may include narrowing the data used in a training set. Specific regions may be targeted for analysis. In embodiments, such targeting can involve an enzyme that either alone, or in combination with other reagent(s), may cleave a DNA sequence or a genome based on its sequence. In some embodiments the enzyme is a restriction enzyme that recognizes and cleaves a specific DNA sequence(s). In other embodiments, more than one restriction enzymes with different recognition sequences can be used in combination. In some embodiments, the restriction enzyme may cleave or not cleave based on the methylation status of the recognition sequences. In some embodiments, the enzyme is one within the CRISPR/Cas family. For example, genomic regions of interest can be targeted using a CRISPR/Cas9 system or other system based on guide RNA (i.e., short RNA sequences which bind to a complementary target DNA sequences and in the process guides an enzyme to act at a target genomic location). In some instances, methylation analysis may be possible without alignment to a reference genome.

I. METHYLATION DETECTION WITH SINGLE MOLECULE, REAL-TIME SEQUENCING

[0197] Embodiments of the present disclosure allow for directly detecting base modifications, without enzymatic or chemical conversion. Kinetic features (e.g., sequence context, IPD, and PW) obtained through single molecule, real-time sequencing can be analyzed with machine learning to develop a model to detect modification or the absence of a modification. Modification levels may be used to determine the origin of DNA molecules or the presence or level of the disorder.

[0198] Using Pacific Biosciences SMRT sequencing as an example of single molecule, real-time sequencing for illustration purposes, a DNA polymerase molecule is positioned at the

bottom of wells that serve as zero-mode waveguides (ZMW). The ZMW is a nanophotonic device for confining light to a small observation volume, which can be a hole whose diameter is very small and does not allow the propagation of light in the wavelength range used for detection such that only emission of optical signals from dye-labeled nucleotide incorporated by the immobilized polymerase are detectable against a low and constant background signal (Eid *et al.*, 2009). The DNA polymerase catalyzes the incorporation of fluorescently labeled nucleotides into complementary nucleic acid strands.

[0199] **FIG. 1** shows an example of molecules carrying base modifications that were sequenced by single-molecule circular consensus sequencing. Molecules 102, 104, and 106 carry base modifications. DNA molecules (e.g., molecule 106) may be ligated with hairpin adapters to form ligated molecule 108. Ligated molecule 108 can then form circularized molecule 110. The circularized molecules can bind to immobilized DNA polymerase and can initiate the DNA synthesis. Molecules not carrying base modifications can also be sequenced.

[0200] **FIG. 2** shows an example of molecules carrying methylated and/or unmethylated CpG sites that were sequenced by single molecule, real-time sequencing. DNA molecules were first ligated with hairpin adapters to form circularized molecules which would bind to immobilized DNA polymerase and to initiate the DNA synthesis. In FIG. 2, DNA molecule 202 is ligated with hairpin adapters to form ligated molecule 204. Ligated molecule 204 then forms circularized molecule 206. The molecules without CpG sites can also be sequenced. Circularized molecule 206 includes an unmethylated CpG site 208, which may still be sequenced.

[0201] Once the DNA synthesis was initialized, fluorescently dye-labeled nucleotides would be incorporated into the newly synthesized strand by the immobilized polymerase on the basis of a circular DNA template, leading to the emission of optical signals. Because the DNA templates were circularized, the entire circular DNA template would go through the polymerase multiple times (i.e. one nucleotide in a DNA template would be sequenced multiple times). A sequence generated from the process, in which all bases in the circularized DNA template entirely passed through the DNA polymerase, is called a subread. One molecule in a ZMW would generate multiple subreads because the polymerase can continue around the entire circular DNA template

multiple times. In one embodiment, a subread may only contain a subset of the sequence, base modifications or other molecular information of the circular DNA template because, in one embodiment, of the existence of sequencing errors.

[0202] As illustrated in **FIG. 3**, the arrival times and durations of the resultant fluorescence pulses would allow one to measure the polymerase kinetics. The interpulse duration (IPD) is a metric for the length of a time period between two emission pulses each of which would be suggestive of an incorporated fluorescently labeled nucleotide in a nascent strand (FIG. 3). As shown in FIG. 3, the pulse width (PW) is another metric, reflecting polymerase kinetics, in association with the duration of the pulses related to a base call. PW could be the duration of the pulse at 0% of the height of signal peak (i.e., fluorescent intensity of dye-labeled nucleotide as incorporated). In one embodiment, PW could be defined by, for example, but not limited to, the duration of the pulse at 5%, 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, or 90% of the height of signal peak. In some embodiments, the PW may be the area under the peak divided by the height of the signal peak.

[0203] Such polymerase kinetics such as IPDs have been shown to be affected by base modifications such as N6-methyladenine (6mA), 5-methylcytosine (5mC) and 5-hydroxymethylcytosine (5hmC) in synthetic and microbial sequences (e.g. *E. coli*) (Flusberg *et al.*, 2010). Flusberg *et al.* 2010 did not use sequence context and IPD as independent inputs to detect a modification, which resulted in a model that lacked practically meaningful accuracy for detection. Flusberg *et al.* only used sequence context to confirm 6mA occurred in GATC. Flusberg *et al.* is silent as to using sequence context in combination with IPD as inputs to detect methylation status.

[0204] The weak interruptions conferred on the new base incorporation to the 5-methylcytosine in complementary strands make methylation calling extremely challenging even for relatively simple microbial genomes when using IPD signals only, as it was reported that the detection of methylation motif $C^{m}CWGG$ was only ranged from 1.9% to 4.3% (Clark *et al.*, 2013). For example, the analytic software package (SMRT Link v6.0.0) provided by Pacific Biosciences is not able to perform 5mC analysis. Furthermore, a previous version of the SMRT

Link v5.1.0 required one to use the Tet1 enzyme to convert 5mC to 5-carboxylcytosine (5caC) prior to methylation analysis, since the IPD signals associated with 5caC would be enhanced (Clark *et al.*, 2013). Thus, it is not surprising that there are no studies showing the feasibility of using single molecule, real-time sequencing to analyze native DNA in a genomewide manner for the human genome.

II. MEASUREMENT WINDOW PATTERNS AND MACHINE LEARNING MODELS

[0205] Techniques to detect modifications in bases without enzymatic or chemically converting the modification and/or the base are desired. As described herein, modifications in a target base may be detected using kinetic feature data obtained from single molecule, real-time sequencing for bases surrounding the target base. Kinetic features may include interpulse duration, pulse width, and sequence context. These kinetic features may be obtained for a measurement window of a certain number of nucleotides upstream and downstream of the target base. These features (e.g., at particular locations in the measurement window) can be used to train a machine learning model. As an example of the sample preparation, the two strands of a DNA molecule may be connected by hairpin adapters, thereby forming a circular DNA molecule. The circular DNA molecule allows for kinetic features to be obtained for either or both of the Watson and Crick strands. A data analysis framework can be developed based on the kinetic features in the measurement windows. This data analysis framework may then be used to detect modifications, including methylation. The section describes various techniques for detecting modifications.

A. Using single strand

[0206] As shown in **FIG. 4**, as an example, we obtained the subreads of the Watson strand from Pacific Biosciences SMRT sequencing to analyze one particular base regarding the states of base modifications. In FIG. 4, the 3 bases from each side of a base that was subjected to base modification analysis would be defined as a measurement window 400. In one embodiment, sequence context, IPDs, and PWs for these 7 bases (i.e. 3-nucleotide (nt) upstream and

downstream sequence and one nucleotide for base modification analysis) were compiled into a 2-dimensional (i.e. 2-D) matrix as a measurement window. In the example shown, the measurement window 400 is for one subread of the Watson strand. Other variations are described herein.

[0207] The first row 402 of the matrix indicated the sequence that was studied. In the second row 404 of the matrix, the position of 0 represented the base for base modification analysis. The relative positions of -1, -2, and -3 indicated the position 1-nt, 2-nt, and 3-nt, respectively, upstream of the base that was subjected to base modification analysis. The relative positions of +1, +2, and +3 indicated the position 1-nt, 2-nt and 3-nt, respectively, downstream of the base that was subjected to base modification analysis. Each position includes 2 columns, which contain the corresponding IPD and PW values. The following 4 rows (rows 408, 412, 416, and 420) corresponded to 4 types of nucleotides (A, C, G, and T) in the strand (e.g. Watson strand), respectively. The presence of IPD and PW values in the matrix depended on which corresponding nucleotide type was sequenced at a particular position. As shown in FIG. 4, at the relative position of 0, the IPD and PW values were shown in the row indicating 'G' in the Watson strand, suggesting that a guanine was called in the sequence result at that position. The other grids in a column that did not correspond to a sequenced base would be coded as '0'. As an example, the sequence information corresponding to the 2-D digital matrix (FIG. 4) would be 5'-GATGACT-3' for the Watson strand.

[0208] As shown in one embodiment depicted in **FIG. 5**, the measurement window could be applied to data from the Crick strand. We obtained the subreads of the Crick strand from single molecule, real-time sequencing to analyze one particular base regarding the states of base modifications. In FIG. 5, the 3 bases from each side of a base that was subjected to base modification analysis and the base subjected to base modification analysis would be defined as a measurement window. In one embodiment, sequence context, IPDs, PWs for these 7 bases (i.e. 3-nucleotide (nt) upstream and downstream sequence and one nucleotide for base modification analysis) were compiled into a 2-dimensional (i.e. 2-D) matrix as a measurement window. The first row of the matrix indicated the sequence that was studied. In the second row of the matrix, the position of 0 represented the base for base modification analysis. The relative positions of -1,

-2, and -3 indicated the position 1-nt, 2-nt and 3-nt, respectively, upstream of the base that was subjected to base modification analysis. The relative positions of +1, +2, and +3 indicated the position 1-nt, 2-nt and 3-nt, respectively, downstream of the base that was subjected to base modification analysis. Each position includes 2 columns, which contained the corresponding IPD and PW values. The following 4 rows corresponded to 4 types of nucleotides (A, C, G, and T) in this strand (e.g. the Crick strand). The presence of IPD and PW values in the matrix depended on which corresponding nucleotide type was sequenced at a particular position. As shown in FIG. 5, at the relative position of 0, the IPD and PW values were shown in the row indicating 'T' in the Crick strand, suggesting that a thymine was called in the sequence result at that position. The other grids in a column that did not correspond to a sequenced base would be coded as '0'. As an example, the sequence information corresponding to the 2-D digital matrix (FIG. 5) would be 5'-ACTTAGC-3' for the Crick strand.

B. Using both Watson and Crick strands

[0209] **FIG. 6** shows an embodiment where the measurement window could be implemented in a way that data from the Watson strand and its complementary Crick strand could be combined. As shown in FIG. 6, we obtained the subreads of the Watson and Crick strands from single molecule, real-time sequencing to analyze one particular base for modifications. In one embodiment, the measurement window from the Crick strand of the circular DNA template was complementary to the measurement window from the Watson strand, which was subjected to base modification analysis. In FIG. 6, the 3 bases from each side of the first base in the Watson strand that was subject to base modification analysis and the first base would be defined as the first measurement window. The 3 bases from each side of the second base in the Crick strand and the second base would be defined as the second measurement window. The second base was complementary to the first base. In one embodiment, the sequence context, IPDs, PWs for these 7 bases (i.e. 3-nucleotide (nt) upstream and downstream sequence and one nucleotide for base modification analysis) from the Watson and Crick strands were compiled into a 2-dimensional (i.e. 2-D) matrices. These measurement windows from the Watson and Crick strands were considered as the first and second measurement windows, respectively.

[0210] The first row of the matrix of the Watson and Crick strands indicated the sequence that was studied. In the second row of the matrix of the Watson strand, the position of 0 represented the first base for base modification analysis. The position of 0 shown in the second row of the matrix of the Crick strand represented the second base complementary to the first base. The relative positions of -1, -2, and -3 indicated the position 1-nt, 2-nt, and 3-nt, respectively, upstream of the first and second bases. The relative positions of +1, +2, and +3 indicated the position 1-nt, 2-nt, and 3-nt, respectively, downstream of the first and second bases. Each position derived from the Watson and Crick strands would correspond to 2 columns which contained the corresponding IPD and PW values. The following 4 rows in the matrices of the Watson and Crick strands corresponded to 4 types of nucleotides (A, C, G, and T) in the specific strand (e.g., the Crick strand), respectively. The presence of IPD and PW values in the matrix depended on which corresponding nucleotide type was sequenced at a particular position.

[0211] As shown in FIG. 6, at the relative position of 0, the IPD and PW values were shown in the row indicating 'A' in the Watson strand and 'T' in the Crick strand, suggesting that an adenine and thymine were called in the sequence result at that position of the Watson and Crick strands, respectively. The other grids in a column that did not correspond to sequenced base would be coded as '0'. As an example, the sequence information corresponding to the 2-D digital matrix of the Watson strand (FIG. 6) would be 5'-ATAAGTT-3'. The sequence information corresponding to the 2-D digital matrix of the Crick strand (FIG. 6) would be 5'-AACTTAT-3'.

[0212] As shown in this example, data from the Watson and Crick strands can be combined to form a new matrix, which may also be considered as a measurement window. This new matrix can be used used as a single sample that is used to train a machine learning model. Thus, all of the values in the new matrix can be treated as separate features, although the particular placement in the 2D matrix can have an impact, e.g., when a convolutional neural network (CNN) is used. The sequence context at the various positions for the different strands can be conveyed via the non-zero entries in the matrix.

[0213] **FIG. 7** shows that the measurement window could be implemented in a way that data from the Watson and Crick strands are not the positions exactly complementary to each other. As

shown in FIG. 7, the first measurement window was 5'-ATAAGTT-3'; and the second measurement window was 5'-GTAACGC-3'. In some embodiments, the Watson and Crick strands may be shifted from each other so that the positions are not complementary.

[0214] **FIG. 8** shows that a measurement window could be used to analyze the methylation states at CpG sites. The position of 0 corresponds to the cytosine of the CpG site, and thus there is a shift by one position between the two strands, so that the C is at the 0 position for both strands. Accordingly, only a portion of the sequences included in the measurement window from the Watson and Crick strands are complementary to each other. In other embodiments, all of the sequences in the measurement window from the Watson and Crick strands can be complementary to each other. In yet other embodiments, none of the sequences in the measurement window from the Watson and Crick strands are complementary to each other.

[0215] In one embodiment, for a measurement window, the length of DNA stretch surrounding a base that was subjected to base modification analysis could be asymmetrical. For example, X-nt upstream and Y-nt downstream of that base could be used for base modification analysis. X could include, but is not limited to, 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 100, 150, 200, 300, 400, 500, 1000, 2000, 4000, 5000, and 10000; Y could include, but was not limited to, 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 100, 150, 200, 300, 400, 500, 1000, 2000, 4000, 5000, and 10000.

C. Training models and detecting modifications

[0216] **FIG. 9** shows a general procedure as to how to use the measurement window to determine any base modifications. DNA samples known to be unmodified and modified were subjected to single molecule, real-time sequencing. The modified DNA (e.g., modified molecule 902) means that the base (e.g., base 904) has the modification (e.g., methylation) at the site. The unmodified DNA (e.g., unmodified molecule 906) means that the base (e.g., base 908) does not have the modification at the site. Both sets of DNA can be artificially created or processed to form the modified/unmodified DNA.