DIAGNOSING CANCER USING GENOMIC SEQUENCING

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 "DETERMINING A NUCLEIC ACID SEQUENCE IMBALANCE," (Attorney Docket No. 016285-005210US) 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 fetal chromosomal aneuploidy by determining imbalances between different nucleic acid sequences, and more particularly to the identification of trisomy 21 (Down syndrome) and other chromosomal aneuploidies via testing a maternal sample (e.g. blood).

BACKGROUND

[0004] Fetal chromosomal aneuploidy results from the presence of abnormal dose(s) of a chromosome or chromosomal region. The abnormal dose(s) can be abnormally high, e.g. the presence of an extra chromosome 21 or chromosomal region in trisomy 21; or abnormally low, e.g. the absence of a copy of chromosome X in Turner syndrome.

[0005] Conventional prenatal diagnostic methods of a fetal chromosomal aneuploidy, e.g., 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 the chromosomal aneuploidy, e.g., 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 of fetal nucleic acids (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 nucleic acid 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 fetal DNA 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.

[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). Digital PCR is an amplification based nucleic acid analysis technique which requires the distribution of a specimen containing nucleic acids into a multitude of discrete samples where each sample containing on average not more than about one target sequence per sample. Specific nucleic acid targets are amplified with sequence-specific primers to generate specific amplicons by digital PCR. The nucleic acid loci to be targeted and the species of or panel of sequence-specific primers to be included in the reactions are determined or selected prior to nucleic acid analysis.

[0011] 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).

[0012] In methods used in the previous studies, the amount of data collected from the digital PCR is quite low. Thus, the accuracy can be compromised due to the small number of data points and typical statistical fluctuations.

[0013] It is therefore desirable that noninvasive tests have high sensitivity and specificity to minimize false negatives and false positives, respectively. However, fetal DNA is present in low absolute concentration and represent a minor portion of all DNA sequences in maternal

plasma and serum. It is therefore also desirable to have methods that 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.

BRIEF SUMMARY

[0013a] In one aspect, there is provided a method of analyzing a biological sample of an organism for deletions or amplifications in one or more chromosomal regions associated with cancer in the organism, the biological sample being obtained from the organism and including cell-free nucleic acid molecules originating from non-malignant cells and potentially from tumor cells associated with cancer, the method comprising:

receiving sequences obtained from a random sequencing of nucleic acid molecules contained in the biological sample;

aligning at least a portion of the sequences to a human genome;

determining a first amount of the sequences obtained from the random sequencing of nucleic acid molecules contained in the biological sample that are identified as aligning to a first chromosomal region that is part of a first chromosome, wherein an aberration in the first chromosomal region is associated with cancer;

determining a second amount of the sequences obtained from the random sequencing of nucleic acid molecules contained in the biological sample that are identified as aligning to one or more second chromosomal regions;

determining a first parameter from the first amount and the second amount, wherein the first parameter includes a ratio between the first and second amounts; and comparing the first parameter to one or more cutoff values to determine a classification of whether the first chromosomal region exhibits a deletion and/or an amplification associated with cancer in the organism.

[0014] Examples of this invention provide methods, systems, and apparatus for determining whether a nucleic acid sequence imbalance (e.g., chromosome imbalance) exists within a biological sample obtained from a pregnant female. This determination may be done by using a parameter of an amount of a clinically-relevant chromosomal region in relation to other non-clinically-relevant chromosomal regions (background regions) within a biological sample. In one example, an amount of chromosomes is determined from a sequencing of nucleic acid molecules in a maternal sample, such as urine, plasma, serum, and other suitable biological samples. Nucleic acid molecules of the biological sample are sequenced, such that a fraction of the genome is sequenced. One or more cutoff values are chosen for determining whether a change compared to a reference quantity exists (i.e. an imbalance), for example, with regards to the ratio of amounts of two chromosomal regions (or sets of regions).

[0015] According to one example, a biological sample received from a pregnant female is analyzed to perform a prenatal diagnosis of a fetal chromosomal aneuploidy. The biological sample includes nucleic acid molecules. A portion of the nucleic acid molecules contained in the biological sample are sequenced. In one example, 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.

[0016] Based on the sequencing, a first amount of a first chromosome is determined from sequences identified as originating from the first chromosome. A second amount of one or more second chromosomes is determined from sequences identified as originating from one of the second chromosomes. A parameter from the first amount and the second amount is then compared to one or more cutoff values. Based on the comparison, a classification of whether a fetal chromosomal aneuploidy exists for the first chromosome is determined. The sequencing advantageously maximizes 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.

[0017] According to one example, a biological sample received from a pregnant female is analyzed to perform a prenatal diagnosis of a fetal chromosomal aneuploidy. The biological sample includes nucleic acid molecules. A percentage of fetal DNA in the biological sample is identified. A number N of sequences to be analyzed based on a desired accuracy is calculated based on the percentage. At least N of the nucleic acid molecules contained in the biological sample are randomly sequenced.

[0018] Based on the random sequencing, a first amount of a first chromosome is determined from sequences identified as originating from the first chromosome. A second amount of one or more second chromosomes is determined from sequences identified as originating from one of the second chromosomes. A parameter from the first amount and the second amount is then compared to one or more cutoff values. Based on the comparison, a classification of whether a fetal chromosomal aneuploidy exists for the first chromosome is determined. The random sequencing advantageously maximizes 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.

[0018a] In another aspect, there is provided a computer program product comprising a computer readable medium encoded with a plurality of instructions for controlling a computing system to perform the method as defined above.

[0019] Other examples of the disclosure are directed to systems and computer readable media associated with methods described herein.

[0020] 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

[0021] FIG. 1 is a flowchart of a method 100 for performing prenatal diagnosis of a fetal chromosomal aneuploidy in a biological sample obtained from a pregnant female subject according to an embodiment of the present invention.

[0022] FIG. 2 is a flowchart of a method 200 for performing prenatal diagnosis of a fetal chromosomal aneuploidy using random sequencing according to an embodiment of the present invention.

[0023] FIG. 3A shows a plot of percentage representation of chromosome 21 sequences in maternal plasma samples involving trisomy 21 or euploid fetuses according to an embodiment of the present invention.

[0024] FIG. 3B shows a correlation between maternal plasma fractional fetal DNA concentrations determined by massively parallel sequencing and microfluidics digital PCR according to an embodiment of the present invention.

[0025] FIG. 4A shows a plot of percentage representation of aligned sequences per chromosome according to an embodiment of the present invention.

[0026] FIG. 4B shows a plot of difference (%) in percentage representation per chromosome between the trisomy 21 case and euploid case shown in FIG. 4A.

[0027] FIG. 5 shows a correlation between degree of over-representation in chromosome 21 sequences and the fractional fetal DNA concentrations in maternal plasma involving trisomy 21 fetuses according to an embodiment of the present invention.

[0028] FIG. 6 shows a table of a portion of human genome that was analyzed according to an embodiment of the present invention. T21 denote a sample obtained from a pregnancy involving a trisomy 21 fetus.

[0029] FIG. 7 shows a table of a number of sequences required to differentiate euploid from trisomy 21 fetuses according to an embodiment of the present invention.

[0030] FIG. 8A shows a table of top ten starting positions of sequenced tags aligned to chromosome 21 according to an embodiment of the present invention.

[0031] FIG. 8B shows a table of top ten starting positions of sequenced tags aligned to chromosome 22 according to an embodiment of the present invention.

[0032] FIG. 9 shows a block diagram of an exemplary computer apparatus usable with system and methods according to embodiments of the present invention.

DEFINITIONS

[0033] 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.

[0034] 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.

[0035] 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).

[0036] 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.

[0037] 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 possible chromosomal 21 imbalance, effectively reducing the need of sample volume to 1/5.

[0038] 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.

[0039] 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.

[0040] 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.

[0041] 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.

[0042] 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.

[0043] 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.

[0044] 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).

[0045] 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.

[0046] The term "*chromosomal aneuploidy*" as used herein means a variation in the quantitative amount of a chromosome from that of a diploid genome. The variation may be a gain or a loss. It may involve the whole of one chromosome or a region of a chromosome.

[0047] The term "*random sequencing*" as used herein refers to sequencing whereby the nucleic acid fragments sequenced have not been specifically identified or targeted before the sequencing procedure. Sequence-specific primers to target specific gene loci are not required. The pools of nucleic acids sequenced vary from sample to sample and even from analysis to analysis for the same sample. The identities of the sequenced nucleic acids are only revealed from the sequencing output generated. In some embodiments of the present invention, the random sequencing may be preceded by procedures to enrich a biological sample with particular populations of nucleic acid molecules sharing certain common features. In one embodiment, each of the fragments in the biological sample have an equal probability of being sequenced.

[0048] The term "*fraction of the human genome*" or "*portion of the human genome*" as used herein refers to less than 100% of the nucleotide sequences in the human genome which comprises of some 3 billion basepairs of nucleotides. In the context of sequencing, it refers to less than 1-fold coverage of the nucleotide sequences in the human genome. The term may be expressed as a percentage or absolute number of nucleotides/basepairs. As an example of use, the term may be used to refer to the actual amount of sequencing performed. Embodiments may determine the required minimal value for the sequenced fraction of the human genome to obtain an accurate diagnosis. As another example of use, the term may

refer to the amount of sequenced data used for deriving a parameter or amount for disease classification.

[0049] The term "*sequenced tag*" as used herein refers to string of nucleotides sequenced from any part or all of a nucleic acid molecule. For example, a sequenced tag may be a short string of nucleotides sequenced from a nucleic acid fragment, a short string of nucleotides at both ends of a nucleic acid fragment, or the sequencing of the entire nucleic acid fragment that exists in the biological sample. A nucleic acid fragment is any part of a larger nucleic acid molecule. A fragment (e.g. a gene) may exist separately (i.e. not connected) to the other parts of the larger nucleic acid molecule.

DETAILED DESCRIPTION

[0050] Embodiments of this invention provide methods, systems, and apparatus for determining whether an increase or decrease (diseased state) of a clinically-relevant chromosomal region exists compared to a non-diseased state. This determination may be done by using a parameter of an amount of a clinically-relevant chromosomal region in relation to other non-clinically-relevant chromosomal regions (background regions) within a biological sample. Nucleic acid molecules of the biological sample are sequenced, such that a fraction of the genome is sequenced, and the amount may be determined from results of the sequencing. One or more cutoff values are chosen for determining whether a change compared to a reference quantity exists (i.e. an imbalance), for example, with regards to the ratio of amounts of two chromosomal regions (or sets of regions).

[0051] 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.

[0052] The clinically relevant chromosomal region (also called a 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. 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.

I. GENERAL METHOD

[0053] FIG. 1 is a flowchart of a method 100 for performing prenatal diagnosis of a fetal chromosomal aneuploidy in a biological sample obtained from a pregnant female subject according to an embodiment of the present invention.

[0054] In step 110, a biological sample from the pregnant female is received. The biological sample may be plasma, urine, serum, or any other suitable sample. The sample contains nucleic acid molecules from the fetus and the pregnant female. For example, the nucleic acid molecules may be fragments from chromosomes.

[0055] In step 120, at least a portion of a plurality of the nucleic acid molecules contained in the biological sample are sequenced. The portion sequenced represents a fraction of the human genome. In one embodiment, the nucleic acid molecules are fragments of respective chromosomes. One end (e.g. 35 basepairs (bp)), both ends, or the entire fragment may be sequenced. All of the nucleic acid molecules in the sample may be sequenced, or just a subset may be sequenced. This subset may be randomly chosen, as will be described in more detail later.

[0056] In one embodiment, the sequencing is done using massively parallel sequencing. 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.

[0057] 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.

[0058] In step 130, based on the sequencing (e.g. data from the sequencing), a first amount of a first chromosome (e.g. the clinically relevant chromosome) is determined. The first amount is determined from sequences identified as originating from the first chromosome. For example, 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.

[0059] In step 140, based on the sequencing, a second amount of one or more second chromosomes is determined from sequences identified as originating from one of the second chromosomes. In one embodiment, the second chromosomes are all of the other chromosomes besides the first one (i.e. the one being tested). In another embodiment, the second chromosome is just a single other chromosome.

[0060] There are a number of ways of determining the amounts of the chromosomes, including but not limited to counting the number of sequenced tags, the number of sequenced nucleotides (basepairs) or the accumulated lengths of sequenced nucleotides (basepairs) originating from particular chromosome(s) or chromosomal regions.

[0061] In another embodiment, rules may be imposed on the results of the sequencing to determine what gets counted. In one aspect, an amount may be obtained based on a proportion of the sequenced output. For example, sequencing output corresponding to

nucleic acid fragments of a specified size range could be selected after the bioinformatics analysis. Examples of the size ranges are about < 300 bp, < 200 bp or <100 bp.

[0062] In step 150, a parameter is determined from the first amount and the second amount. The parameter may be, for example, a simple ratio of the first amount to the second amount, or the first amount to the second amount plus the first amount. In one aspect, each amount could be an argument to a function or separate functions, where a ratio may be then taken of these separate functions. One skilled in the art will appreciate the number of different suitable parameters.

[0063] 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 including the clinically-relevant chromosome) or a particular subset of chromosomes (e.g. just one other chromosome than the one being tested.)

[0064] In step 150, the parameter is compared to one or more cutoff values. The cutoff values may be determined from any number of suitable ways. Such ways include Bayesian-type likelihood method, sequential probability ratio testing (SPRT), false discovery, confidence interval, receiver operating characteristic (ROC). Examples of applications of these methods and sample-specific methods are described in concurrently filed application "DETERMINING A NUCLEIC ACID SEQUENCE IMBALANCE," (Attorney Docket No. 016285-005210US), which is incorporated by reference.

[0065] In one embodiment, the parameter (e.g. the fractional representation of the clinically relevant chromosome) 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 (i.e. the cutoff values) 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.

[0066] In step 160, based on the comparison, a classification of whether a fetal chromosomal aneuploidy exists for the first chromosome is determined. In one embodiment, the classification is a definitive yes or no. In another embodiment, a classification may be unclassifiable or uncertain. In yet another embodiment, the classification may be a score that is to be interpreted at a later date, for example, by a doctor.

II. SEQUENCING, ALIGNING, AND DETERMINING AMOUNTS

[0067] As mentioned above, only a fraction of the genome is sequenced. In one aspect, even when a pool of nucleic acids in a specimen is sequenced at <100% genomic coverage instead of at several folds of coverage, and among the proportion of captured nucleic acid molecules, most of each nucleic acid species is only sequenced once. Also, dosage imbalance of a particular chromosome or chromosomal regions can be quantitatively determined. In other words, the dosage imbalance of the chromosome or chromosomal regions is inferred from the percentage representation of the said locus among other mappable sequenced tags of the specimen.

[0068] This is contrasted from situations where the same pool of nucleic acids is sequenced multiple times to achieve high redundancy or several folds of coverage whereby each nucleic acid species is sequenced multiple times. In such situations, the number of times a particular nucleic acid species have been sequenced relative to that of another nucleic acid species correlate with their relative concentrations in the original sample. The sequencing cost increases with the number of fold coverage required to achieve accurate representation of the nucleic acid species.

[0069] 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.

[0070] 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 sequence 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.

[0071] 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. As paired end sequencing allows one to deduce the size of the original nucleic acid fragment, one example is to focus on the counting of the number of paired sequenced tags corresponding to nucleic acid fragments of a specified size, such as < 300 bp, < 200bp or < 100 bp.

[0072] 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, as discussed above, 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). In one embodiment, a portion or subset of the pre-selected pool of nucleic acids is sequenced randomly.

[0073] 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.

III. DETERMINING AMOUNTS OF CHROMOSOMES FROM SEQUENCING OUTPUT

[0074] After the massively parallel sequencing, bioinformatics analysis was performed to locate the chromosomal origin of the sequenced tags. After this procedure, tags identified as originating from the potentially aneuploid chromosome, i.e. chromosome 21 in this study, are compared quantitatively to all of the sequenced tags or tags originating from one of more chromosomes not involved in the aneuploidy. The relationship between the sequencing output from chromosome 21 and other non-21 chromosomes for a test specimen is compared with cut-off values derived with methods described in the above section to determine if the specimen was obtained from a pregnancy involving a euploid or trisomy 21 fetus.

[0075] A number of different amounts include but not limited to the following could be derived from the sequenced tags. For example, the number of sequenced tags, i.e. absolute count, aligned to a particular chromosome could be compared to the absolute count of sequenced tags aligned to other chromosomes. Alternatively, the fractional count of the amount of sequenced tags from chromosome 21 with reference to all or some other sequenced tags could be compared to that of other non-aneuploid chromosomes. In the present experiment, because 36 bp were sequenced from each DNA fragment, the number of nucleotides sequenced from a particular chromosome could easily be derived from 36 bp multiplied by the sequenced tag count.

[0076] Furthermore, as each maternal plasma specimen was only sequenced using one flow cell which could only sequence a fraction of the human genome, by statistics, most of the maternal plasma DNA fragment species would only each have been sequenced to generate one sequenced tag count. In other words, the nucleic acid fragments present in the maternal

plasma specimen were sequenced at less than 1-fold coverage. Thus, the total number of sequenced nucleotides for any particular chromosome would mostly correspond to the amount, proportion or length of the part of the said chromosome that has been sequenced. Hence, the quantitative determination of the representation of the potentially aneuploid chromosome could be derived from a fraction of the number or equivalent length of nucleotides sequenced from that chromosome with reference to a similarly derived quantity for other chromosomes.

IV. ENRICHMENT FOR POOLS OF NUCLEIC ACIDS FOR SEQUENCING

[0077] As mentioned above and established in the example section below, only a portion of the human genome needs to be sequenced to differentiate trisomy 21 from euploid cases. Thus, it would be possible and cost-effective to enrich the pool of nucleic acids to be sequenced prior to random sequencing of a fraction of the enriched pool. For example, 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.

[0078] 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). The proportion of fetal derived sequences would be enriched in the nucleic acid pool comprised of shorter fragments. According to FIG. 7, the number of sequenced tags required for differentiating euploid from trisomy 21 cases would reduce as the fractional fetal DNA concentration increases.

[0079] Alternatively, sequences originating from a potentially aneuploid chromosome and one or more chromosomes not involved in the aneuploidy could be enriched by hybridization techniques for example onto oligonucelotide microarrays. The enriched pools of nucleic acids would then be subjected to random sequencing. This would allow the reduction in sequencing costs.

V. RANDOM SEQUENCING

[0080] FIG. 2 is a flowchart of a method 200 for performing prenatal diagnosis of a fetal chromosomal aneuploidy using random sequencing according to an embodiment of the present invention. In one aspect for the massively parallel 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.

[0081] In step 210, a biological sample from the pregnant female is received. In step 220, the number N of sequences to be analyzed is calculated for a desired accuracy. In one embodiment, a percentage of fetal DNA in the biological sample is first identified. This may be done by any suitable means as will be known to one skilled in the art. The identification may simply be reading a value that was measured by another entity. In this embodiment, the calculation of the number N of sequences to be analyzed is based on the percentage. For example, the number of sequences needed to be analyzed would be increased when the fetal DNA percentage drops, and could be decreased when the fetal DNA rises. The number N may be a fixed number or a relative number, such as a percentage. In another embodiment, one could sequence a number N that is known to be adequate for accurate disease diagnosis. The number N could be made sufficient even in pregnancies with fetal DNA concentrations that are at the lower end of the normal range.

[0082] In step 230, at least N of a plurality of the nucleic acid molecules contained in the biological sample are randomly sequenced. A feature of this described approach is that the nucleic acids to be sequenced are not specifically identified or targeted before sample analysis, i.e. sequencing. Sequence-specific primers to target specific gene loci are not needed for sequencing. The pools of nucleic acids sequenced vary from sample to sample and even from analysis to analysis for the same sample. Furthermore, from the below descriptions (FIG. 6), the amount of sequencing output required for case diagnosis could vary between the tested specimens and the reference population. These aspects are in marked contrast to most molecular diagnostic approaches, such as those based on fluorescence in situ hybridization, quantitative florescence PCR, quantitative real-time PCR, digital PCR, comparative genomic hybridization, microarray comparative genomic hybridization and so on, where gene loci to

be targeted require prior pre-determination, thus requiring the use of locus-specific primers or probe sets or panels of such.

[0083] 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%, 1%, 5%, 10%, 20%, or 30% of the genome. In other embodiments, the fraction is at least any one of these values.

[0084] The rest of the steps 240-270 may proceed in a similar manner as method 100.

VI. POST-SEQUENCING SELECTION OF POOLS OF SEQUENCED TAGS

[0085] As described in examples II and III below, a subset of the sequenced data is sufficient to distinguish trisomy 21 from euploid cases. The subset of sequenced data could be the proportion of sequenced tags that passed certain quality parameters. For example, in example II, sequenced tags that were uniquely aligned to the repeat-masked reference human genome were used. Alternatively, one may sequence a representative pool of nucleic acid fragments from all of the chromosomes but focus on the comparison between data relevant to the potentially aneuploid chromosome and data relevant to a number of non-aneuploid chromosomes.

[0086] Yet alternatively, a subset of the sequencing output encompassing sequenced tags generated from nucleic acid fragments corresponding to a specified size window in the original specimen could be sub-selected during the post-sequencing analysis. For example, using the Illumina Genome analyzer, one could use paired-end sequencing which refers to sequencing the two ends of nucleic acid fragments. The sequenced data from each paired-end are then aligned to the reference human genome sequence. The distance or number of nucleotides spanning between the two ends could then be deduced. The whole length of the original nucleic acid fragment could also be deduced. Alternatively, sequencing platforms such as the 454 platform and possibly some single molecule sequencing techniques are able to sequence the full length of short nucleic acid fragments, for example 200 bp. In this

manner, the actual length of the nucleic acid fragment would be immediately known from the sequenced data.

[0087] Such paired-end analysis is also possible using other sequencing platforms, e.g. the Applied Biosystems SOLiD system. For the Roche 454 platform, because of its increased read length compared with other massively parallel sequencing systems, it is also possible to determine the length of a fragment from its complete sequence.

[0088] The advantage of focusing the data analysis on the subset of sequenced tags corresponding to short nucleic acid fragments in the original maternal plasma specimen because the dataset would effectively be enriched with DNA sequences derived from the fetus. This is because the 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). According to FIG. 7, the number of sequenced tags required for differentiating euploid from trisomy 21 cases would reduce as the fractional fetal DNA concentration increases.

[0089] The post-sequencing selection of subsets of nucleic acid pools is different from other nucleic acid enrichment strategies which are performed prior to specimen analysis, such as the use gel electrophoresis or size exclusion columns for the selection of nucleic acids of particular sizes, which require the physical separation of the enriched pool from the background pool of nucleic acids. The physical procedures would introduce more experimental steps and may be prone to problems such as contamination. The post-sequencing *in silico* selection of subsets of sequencing output would also allow one to vary the selection depending on the sensitivity and specificity required for disease determination.

[0090] The bioinformatics, computational and statistical approaches used to determine if a maternal plasma specimen is obtained from a pregnant woman conceived with a trisomy 21 or euploid fetus could be compiled into a computer program product used to determine parameters from the sequencing output. The operation of the computer program would involve the determining of a quantitative amount from the potentially aneuploid chromosome as well as amount(s) from one or more of the other chromosomes. A parameter would be determined and compared with appropriate cut-off values to determine if a fetal chromosomal aneuploidy exists for the potentially aneuploid chromosome.

EXAMPLES

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

I. PRENATAL DIAGNOSIS OF FETAL TRISOMY 21

[0091] Eight pregnant women were recruited for the study. All of the pregnant women were in the 1st or 2nd trimester of gestation and had a singleton pregnancy. Four of them were each carrying a fetus with trisomy 21 and the other four were each carrying a euploid fetus. Twenty milliliters of peripheral venous blood was collected from each subject. Maternal plasma was harvested after centrifugation at 1600 x g for 10 minutes and further centrifuged at 16000 x g for 10 minutes. DNA was then extracted from 5-10 mL of each plasma sample. The maternal plasma DNA was then used for massively parallel sequencing by the Illumina Genome Analyzer according to manufacturer's instructions. The technicians performing the sequencing were blinded from the fetal diagnoses during the sequencing and sequence data analysis.

[0092] Briefly, approximately 50 ng of maternal plasma DNA was used for DNA library preparation. It is possible to start with lesser amounts such as 15 ng or 10 ng of maternal plasma DNA. Maternal plasma DNA fragments were blunt-ended, ligated to Solexa adaptors and fragments of 150-300 bp were selected by gel purification. Alternatively, blunt-ended and adaptor-ligated maternal plasma DNA fragments could be passed through columns (e.g. AMPure, Agencourt) to remove unligated adaptors without size-selection before cluster generation. The adaptor-ligated DNA was hybridized to the surface of flow cells, and DNA clusters were generated using the Illumina cluster station, followed by 36 cycles of sequencing on the Illumina Genome Analyzer. DNA from each maternal plasma specimen was sequenced by one flow cell. Sequenced reads were compiled using Solexa Analysis Pipeline. All reads were then aligned to the repeat-masked reference human genomic sequence, NCBI 36 assembly (GenBank accession numbers: NC_000001 to NC_000024), using the Eland application.

[0093] In this study, to reduce the complexity of the data analysis, only sequences that have been mapped to a unique location in the repeat-masked human genome reference are further considered. Other subsets of or the entire set of the sequenced data could alternatively be used. The total number of uniquely mappable sequences for each specimen was counted. The number of sequences uniquely aligned to chromosome 21 was expressed as a proportion to the total count of aligned sequences for each specimen. As maternal plasma contains fetal

DNA among a background of DNA of maternal origin, the trisomy 21 fetus would contribute extra sequenced tags originating from chromosome 21 due to the presence of an extra copy of chromosome 21 in the fetal genome. Hence, the percentage of chromosome 21 sequences in maternal plasma from a pregnancy carrying a trisomy 21 fetus would be higher than that from a pregnancy with a euploid fetus. The analysis does not require the targeting of fetal-specific sequences. It also does not require the prior physical separation of fetal from maternal nucleic acids. It also does not require the need to distinguish or identify fetal from maternal sequences after sequencing.

[0094] FIG. 3A shows the percentage of sequences mapped to chromosome 21 (percentage representation of chromosome 21) for each of the 8 maternal plasma DNA samples. The percentage representation of chromosome 21 was significantly higher in maternal plasma of trisomy 21 pregnancies than in that of euploid pregnancies. These data suggest that noninvasive prenatal diagnosis of fetal aneuploidy could be achieved by determining the percentage representation of the aneuploid chromosome compared to that of a reference population. Alternatively, the chromosome 21 over-representation could be detected by comparing the percentage representation of chromosome 21 obtained experimentally with the percentage representation of chromosome 21 sequences expected for a euploid human genome. This could be done by masking or not masking the repeat regions in the human genome.

[0095] Five of the eight pregnant women were each carrying a male fetus. The sequences mapped to the Y chromosome would be fetal-specific. The percentage of sequences mapped to the Y-chromosome was used to calculate the fractional fetal DNA concentration in the original maternal plasma specimen. Moreover, the fractional fetal DNA concentration was also determined by using microfluidics digital PCR involving the *zinc finger protein, X-linked* (*ZFX*) and *zinc finger protein, Y-linked* (*ZFY*) paralogous genes.

[0096] FIG. 3B shows the correlation of the fractional fetal DNA concentrations as inferred by the percentage representation of Y chromosome by sequencing and that determined by *ZFY/ZFX* microfluidics digital PCR. There was a positive correlation between the fractional fetal DNA concentrations in maternal plasma determined by these two methods. The coefficient of correlation (r) was 0.917 in the Pearson correlation analysis.

[0097] The percentages of maternal plasma DNA sequences aligned to each of the 24 chromosomes (22 autosomes and X and Y chromosomes) for two representative cases are

shown in FIG. 4A. One pregnant woman was carrying a trisomy 21 fetus and the other was carrying a euploid fetus. The percentage representation of sequences mapped to chromosome 21 is higher in the pregnant woman carrying a trisomy 21 fetus when compared with the pregnant woman carrying a normal fetus.

[0098] The differences (%) of the percentage representation per chromosome between the maternal plasma DNA specimens of the above two cases is shown in FIG. 4B. The percentage difference for a particular chromosome is calculated using the formula below:

Percentage difference (%) = $(P_{21} - P_E)/ P_E \times 100\%$, where

P_{21} = percentage of plasma DNA sequences aligned to the particular chromosome in the pregnant woman carrying a trisomy 21 fetus and;

P_E = percentage of plasma DNA sequences aligned to the particular chromosome in the pregnant woman carrying a euploid fetus.

[0099] As shown in FIG. 4B, there is an over-representation of chromosome 21 sequences by 11% in the plasma of the pregnant woman carrying a trisomy 21 fetus when compared with the pregnant woman carrying a euploid fetus. For the sequences aligned to other chromosomes, the differences between the two cases were within 5%. As the percentage representation for chromosome 21 is increased in the trisomy 21 compared with the euploid maternal plasma samples, the difference (%) could be alternatively referred as the degree of over-representation in chromosome 21 sequences. In addition to differences (%) and absolute differences between the chromosome 21 percentage representation, ratios of the counts from test and reference samples could also be calculated and would be indicative of the degree of chromosome 21 over-representation in trisomy 21 compared with euploid samples.

[0100] For the four pregnant women each carrying a euploid fetus, a mean of 1.345% of their plasma DNA sequences were aligned to chromosome 21. In the four pregnant women carrying a trisomy 21 fetus, three of their fetuses were males. The percentage representation of chromosome 21 was calculated for each of these three cases. The difference (%) in chromosome 21 percentage representation for each of these three trisomy 21 cases from the mean chromosome 21 percentage representation derived from values of the four euploid cases were determined as described above. In other words, the mean of the four cases carrying a euploid fetus was used as the reference in this calculation. The fractional fetal

DNA concentrations for these three male trisomy 21 cases were inferred from their respective percentage representation of Y chromosome sequences.

[0101] The correlation between the degree of over-representation for chromosome 21 sequences and the fractional fetal DNA concentrations is shown in FIG. 5. There was a significant positive correlation between the two parameters. The coefficient of correlation (r) was 0.898 in the Pearson correlation analysis. These results indicate that the degree of over-representation of chromosome 21 sequences in maternal plasma is related to the fractional concentration of fetal DNA in the maternal plasma sample. Thus, cut-off values in the degree of chromosome 21 sequence over-representation relevant to the fractional fetal DNA concentrations could be determined to identify pregnancies involving trisomy 21 fetuses.

[0102] 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. For example, we have demonstrated in FIG. 3B that there is good correlation between the fetal DNA concentrations estimated based on the Y-chromosome count generated during the sequencing run and the *ZFY/ZFX* ratio generated external to the sequencing run. 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.

[0103] 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.

[0104] 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.

[0105] 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 sequenced tags aligned to various chromosomal regions is compared between specimens containing or not containing the DNA species of interest. Chromosomal aberrations would be revealed by differences in the number (or percentage) of sequences aligned to any given chromosomal region in the specimens.

[0106] 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.

II. SEQUENCING JUST A FRACTION OF THE HUMAN GENOME

[0107] In the experiment described in example I above, maternal plasma DNA from each individual specimen was sequenced using one flow cell only. The number of sequenced tags generated from each of the tested specimens by the sequencing run is shown in FIG. 6. T21 denote a sample obtained from a pregnancy involving a trisomy 21 fetus.

[0108] As 36 bp were sequenced from each of the sequenced maternal plasma DNA fragments, the number of nucleotides/basepairs sequenced from each specimen could be determined by 36 bp multiplied by the sequenced tag count and are also shown in FIG. 6. As

there are approximately 3 billion basepairs in the human genome, the amount of sequencing data generated from each maternal plasma specimen represented only a fraction, ranging from some 10% to 13%.

[0109] Furthermore, in this study, only the uniquely mappable sequenced tags, termed U0 in nomenclature from the Eland software, were used to demonstrate the presence of over-representation in the amount of chromosome 21 sequences in the maternal plasma specimens from pregnancies each carrying a fetus with trisomy 21, as described in example I above. As shown in FIG. 6, U0 sequences only represent a subset of all the sequenced tags generated from each specimen and further represent an even smaller proportion, some 2%, of the human genome. These data indicate that the sequencing of only a portion of the human genomic sequences present in the tested specimen is sufficient to achieve the diagnosis of fetal aneuploidy.

III. DETERMINATION OF NUMBER OF SEQUENCES REQUIRED

[0110] The sequencing result of the plasma DNA from a pregnant woman carrying a euploid male fetus is used for this analysis. The number of sequenced tags that can be mapped without mismatches to the reference human genome sequence was 1,990,000. Subsets of sequences were randomly chosen from these 1,990,000 tags and the percentage of sequences aligned to chromosome 21 was calculated within each subset. The number of sequences in the subsets was varied from 60,000 to 540,000 sequences. For each subset size, multiple subsets of the same number of sequenced tags were compiled by random selection of the sequenced tags from the total pool until no other combination was possible. The mean percentage of sequences aligned to chromosome 21 and its standard deviation (SD) were then calculated from the multiple subsets within each subset size. These data were compared across different subset sizes to determine the effect of subset size on the distribution of the percentage of sequences aligned to the chromosome 21. The 5^{th} and 95^{th} percentiles of the percentages were then calculated according to the mean and SD.

[0111] When a pregnant woman is carrying a trisomy 21 fetus, the sequenced tags aligned to chromosome 21 should be over-represented in the maternal plasma due to an extra dose of chromosome 21 from the fetus. The degree of over-representation is dependent on the fetal DNA percentage in the maternal plasma DNA sample following the equation below:

$$Per_{T21} = Per_{Eu} \times (1 + f/2)$$

where

Per_{T21} represents the percentage of sequences aligned to chromosome 21 in a woman with a trisomy 21 fetus; and

Per_{Eu} represents the percentage of sequences aligned to chromosome 21 in a woman with a euploid fetus; and

f represents the fetal DNA percentage in maternal plasma DNA

[0112] As shown in FIG. 7, the SD for the percentages of sequences aligned to chromosome 21 decreases with increasing number of sequences in each subset. Therefore, when the number of sequences in each subset increases, the interval between the 5^{th} and 95^{th} percentiles decreases. When the 5%-95% interval for the euploid and trisomy 21 cases do not overlap, then the differentiation between the two groups of cases would be possible with an accuracy of >95%.

[0113] As shown in FIG. 7, the minimal subset size for the differentiation of trisomy 21 cases from euploid cases is dependent on the fetal DNA percentage. The minimal subset sizes for differentiating trisomy 21 from euploid cases were 120,000, 180,000 and 540,000 sequences for fetal DNA percentages of 20%, 10% and 5%, respectively. In other words, the number of sequences needed to be analyzed would be 120,000 for determining whether a fetus has trisomy 21 when a maternal plasma DNA sample contains 20% fetal DNA. The number of sequences needed to be analyzed would be increased to 540,000 when the fetal DNA percentage drops to 5%.

[0114] As the data were generated using 36 basepair sequencing, 120,000, 180,000 and 540,000 sequences correspond to 0.14%, 0.22% and 0.65% of the human genome, respectively. As the lower range of fetal DNA concentrations in maternal plasma obtained from early pregnancies were reported to be some 5% (Lo, YMD et al. 1998 *Am J Hum Genet* 62, 768-775), the sequencing of about 0.6% of the human genome may represent the minimal amount of sequencing required for diagnosis with at least 95% accuracy in detecting fetal chromosomal aneuploidy for any pregnancy.

IV. RANDOM SEQUENCING

[0115] To illustrate that the sequenced DNA fragments were randomly selected during the sequencing run, we obtained the sequenced tags generated from the eight maternal plasma samples analyzed in example I. For each maternal plasma specimen, we determined the starting positions in relation to the reference human genome sequence, NCBI assembly 36, of each of the 36 bp sequenced tags that were aligned uniquely to chromosome 21 without

mismatches. We then ordered the starting position number for the pools of aligned sequenced tags from each specimen in ascending order. We performed a similar analysis for chromosome 22. For illustrative purpose, the top ten starting positions for chromosome 21 and chromosome 22 for each of the maternal plasma specimens are shown in FIGS. 8A and 8B, respectively. As can be appreciated from these Tables, the sequenced pools of DNA fragments were non-identical between samples.

[0116] 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.

[0117] 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.

[0118] An example of a computer system is shown in FIG. 9. The subsystems shown in FIG. 9 are interconnected via a system bus 975. Additional subsystems such as a printer 974, keyboard 978, fixed disk 979, monitor 976, which is coupled to display adapter 982, and others are shown. Peripherals and input/output (I/O) devices, which couple to I/O controller 971, can be connected to the computer system by any number of means known in the art, such as serial port 977. For example, serial port 977 or external interface 981 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 973 to communicate with each subsystem and to control the execution of instructions from system memory 972 or the fixed disk 979, as well as the exchange of information between subsystems. The system memory 972 and/or the fixed disk 979 may embody a computer readable medium.

[0119] 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.

[0120] All publications, patents, and patent applications cited herein are hereby incorporated by reference in their entirety for all purposes.

CLAIMS

1. A method of analyzing a biological sample of an organism for deletions or amplifications in one or more chromosomal regions associated with cancer in the organism, the biological sample being obtained from the organism and including cell-free nucleic acid molecules originating from non-malignant cells and potentially from tumor cells associated with cancer, the method comprising:

receiving sequences obtained from a random sequencing of nucleic acid molecules contained in the biological sample;

aligning at least a portion of the sequences to a human genome;

determining a first amount of the sequences obtained from the random sequencing of nucleic acid molecules contained in the biological sample that are identified as aligning to a first chromosomal region that is part of a first chromosome, wherein an aberration in the first chromosomal region is associated with cancer;

determining a second amount of the sequences obtained from the random sequencing of nucleic acid molecules contained in the biological sample that are identified as aligning to one or more second chromosomal regions;

determining a first parameter from the first amount and the second amount, wherein the first parameter includes a ratio between the first and second amounts; and

comparing the first parameter to one or more cutoff values to determine a classification of whether the first chromosomal region exhibits a deletion and/or an amplification associated with cancer in the organism.

2. The method of claim 1, wherein the classification includes whether the first chromosomal region is amplified, and wherein a classification of amplification is made when the first parameter is greater than a first cutoff value of the one or more cutoff values.

3. The method of claim 1, wherein the classification includes whether the first chromosomal region has a deletion, and wherein a classification of deletion is made when the first parameter is less than a first cutoff value of the one or more cutoff values.

4. The method of claim 1, further comprising:

determining a third amount of the sequences identified as aligning to a third chromosomal region, wherein an aberration in the third chromosomal region is associated with cancer;

determining a second parameter for the third chromosomal region, wherein the second parameter represents a relative amount between the third and second amounts;

comparing the second parameter to the one or more cutoff values to determine a classification of whether the third chromosomal region exhibits a deletion and/or an amplification associated with cancer; and

using the classifications of the first and third chromosomal regions to detect whether the organism has cancer corresponding to the classifications of the first and third chromosomal regions.

5. The method of claim 1, wherein the biological sample is plasma, serum, urine or saliva.

6. The method of claim 1, wherein the biological sample is transcervical lavage fluid.

7. The method of claim 1, wherein the ratio provides a percentage of the sequences identified as aligning to the first chromosomal region among the sequences.

8. The method of claim 1, wherein the ratio is obtained from any one or more of a fractional count of the number of aligned sequences, a fractional number of sequenced nucleotides, and a fractional length of accumulated sequences.

9. The method of claim 1, wherein the sequences that are used to determine the first amount are selected to be less than a specified number of bases.

10. The method of claim 9, wherein the specified number of bases is 300 bases, 200 bases, or 100 bases.

11. The method of claim 1, wherein the nucleic acid molecules of the biological sample have been enriched for sequences originating from the first chromosomal region.

12. The method of claim 1, wherein the nucleic acid molecules of the biological sample have been enriched for sequences less than 300 bases.

13. The method of claim 1, wherein the nucleic acid molecules of the biological sample have been enriched for sequences less than 200 bases.

14. The method of claim 1, wherein the nucleic acid molecules of the biological sample have been amplified using a polymerase chain reaction.

15. The method of claim 1, wherein the aligned portion represents at least a pre-determined fraction of the human genome.

16. The method of claim 1, wherein the pre-determined fraction represents at least 0.1% of the human genome.

17. The method of claim 1, wherein the pre-determined fraction represents at least 0.5% of the human genome.

18. The method of claim 1, wherein at least one of the one or more cutoff values is derived from a percentage of tumor cells in the biological sample.

19. The method of claim 1, wherein at least one of the one or more cutoff values is a reference value established in a normal biological sample.

20. The method of claim 19, wherein the at least one cutoff value is determined from a mean of the value of the first parameter in a plurality of normal biological samples.

21. The method of claim 1, wherein the one or more cutoff values are derived from a normal value for the ratio between the first and second amounts.

22. The method of claim 1, further comprising:

identifying an amount of tumor DNA in the biological sample; and

calculating a number N of sequences to be analyzed based on a desired accuracy.

23. The method of any one of claims 1 to 22, wherein the sequences identified as aligning to the first chromosomal region uniquely align to the first chromosomal region.

24. The method of any one of claims 1 to 23, wherein the sequences identified as aligning to the one or more of the second chromosomal regions uniquely align to only one of the second chromosomal regions.

25. The method of any one of claims 1 to 24, wherein the human genome is repeat-masked.

26. The method of any one of claims 1 to 25, wherein the first amount is determined based on a pool of sequences that align to a plurality of positions of the first chromosomal region.

27. The method of claim 26, wherein which part of the human genome that the sequences are aligned is not pre-determined.

28. The method of claim 27, wherein the part of the human genome corresponding to the sequences aligned to the first chromosomal region is not pre-determined.

29. The method of claim 1, wherein the first parameter is a fractional representation of sequences identified as aligning to the first chromosomal region, the fractional representation being a measure of a proportion of nucleic acid molecules in the biological sample that are from the first chromosomal region.

30. The method of claim 29, wherein the one or more cutoff values take into account a size of the first chromosomal region relative to a size of the one or more second chromosomal regions.

31. The method of claim 1, wherein the sequencing includes paired-end sequencing, and wherein aligning each sequence includes aligning paired-end sequences.

32. The method of claim 1, wherein the first parameter includes a ratio between the first and second amounts by including a simple ratio of the first amount divided by a term including the second amount or by including a ratio of separate functions of the first amount and the second amount.

33. The method of claim 1, further comprising:

performing the random sequencing of nucleic acid molecules contained in the biological sample to obtain the sequences.

34. A computer program product comprising a computer readable medium encoded with a plurality of instructions for controlling a computing system to perform the method of any one of claims 1 to 32.

DIAGNOSING CANCER USING GENOMIC SEQUENCING

Abstract

Embodiments of this invention provide methods, systems, and apparatus for determining whether a fetal chromosomal aneuploidy exists from a biological sample obtained from a pregnant female. Nucleic acid molecules of the biological sample are sequenced, such that a fraction of the genome is sequenced. Respective amounts of a clinically-relevant chromosome and of background chromosomes are determined from results of the sequencing. A parameter derived from these amounts (e.g. a ratio) is compared to one or more cutoff values, thereby determining a classification of whether a fetal chromosomal aneuploidy exists.

Fig. 1