METHOD FOR DETERMINING THE PRESENCE OR ABSENCE OF DIFFERENT ANEUPLOIDIES IN A SAMPLE

FIELD OF THE INVENTION

The invention relates generally to the field of diagnostics, and provides a method for determining variations in the amount of nucleic acid sequences in a mixture of nucleic acids derived from different genomes. In particular, the method is applicable to the practice of noninvasive prenatal diagnostics, and to the diagnosis and monitoring of metastatic progression in cancer patients.

BACKGROUND OF THE INVENTION

One of the critical endeavors in human medical research is the discovery of genetic abnormalities that are central to adverse health consequences. In many cases, specific genes and/or critical diagnostic markers have been identified in portions of the genome that are present at abnormal copy numbers. For example, in prenatal diagnosis, extra or missing copies of whole chromosomes are the frequently occurring genetic lesions. In cancer, deletion or multiplication of copies of whole chromosomes or chromosomal segments, and higher level amplifications of specific regions of the genome, are common occurrences.

Most information about copy number variation has been provided by cytogenetic resolution that has permitted recognition of structural abnormalities. Conventional procedures for genetic screening and biological dosimetry have utilized invasive procedures *e.g.* amniocentesis, to obtain cells for the analysis of karyotypes. Recognizing the need for more rapid testing methods that do not require cell culture, fluorescence in situ hybridization (FISH), quantitative fluorescence PCR (QF-PCR) and array- Comparative Genomic Hybridization (array-CGH) have been developed as molecular-cytogenetic methods for the analysis of copy number variations.

The advent of technologies that allow for sequencing entire genomes in relatively short time, and the discovery of circulating cell-free DNA (cfDNA) have provided the opportunity to compare genetic material originating from one chromosome to be compared to that of another without the risks associated with invasive sampling methods. However, the limitations of the existing methods, which include insufficient sensitivity stemming from the limited levels of cfDNA, and the sequencing bias of the technology stemming from the inherent nature of genomic information, underlie the continuing need for noninvasive methods that would provide any or all of the specificity, sensitivity, and applicability, to reliably diagnose copy number changes in a variety of clinical settings.

The present invention fulfills some of the above needs and in particular offers an advantage in providing a reliable method that is applicable at least to the practice of noninvasive prenatal diagnostics, and to the diagnosis and monitoring of metastatic progression in cancer patients.

SUMMARY OF THE INVENTION

The invention provides a method for determining copy number variations (CNV) of a sequence of interest in a test sample that comprises a mixture of nucleic acids that are known or are suspected to differ in the amount of one or more sequence of interest. The method comprises a statistical approach that accounts for accrued variability stemming from process-related, interchromosomal and inter-sequencing variability. The method is applicable to determining CNV of any fetal aneuploidy, and CNVs known or suspected to be associated with a variety of medical conditions. CNV that can be determined according to the present method include trisomies and monosomies of any one or more of chromosomes 1-22, X and Y, other chromosomal polysomies, and deletions and/or duplications of segments of any one or more of the chromosomes, which can be detected by sequencing only once the nucleic acids of a test sample. Any aneuploidy can be determined from sequencing information that is obtained by sequencing only once the nucleic acids of a test sample.

In one embodiment, a method is provided for determining the presence or absence of any four or more different complete fetal chromosomal aneuploidies in a maternal test sample comprising fetal and maternal nucleic acid molecules. The steps of the method comprise (a) obtaining sequence information for the fetal and maternal nucleic acids in the maternal test sample; (b) using the sequence information to identify a number of sequence tags for each of any four or more chromosomes of interest selected from chromosomes 1-22, X and Y and to identify a number of sequence tags for a normalizing chromosome sequence for each of the any four or more chromosomes of interest; (c) using the number of sequence tags identified for each of the any four or more chromosomes of interest and the number of sequence tags identified for each normalizing chromosome to calculate a single chromosome dose for each of the any four or more chromosomes of interest; and (d) comparing each of the single chromosome doses for each of the any four or more chromosomes of interest to a threshold value for each of the four or more chromosomes of interest, and thereby determining the presence or absence of any four or more complete different fetal chromosomal aneuploidies in the maternal test sample. Step (a) can comprise sequencing at least a portion of the nucleic acid molecules of a test sample to obtain said sequence information for the fetal and maternal nucleic acid molecules of the test sample. In some embodiments, step (c) comprises calculating a single chromosome dose for each of the chromosomes of interest as the ratio of the number of sequence tags identified for each of the chromosomes of interest and the number of sequence tags identified for the normalizing chromosome sequence for each of the chromosomes of interest. In some other embodiments, step (c) comprises (i) calculating a sequence tag density ratio for each of the chromosomes of interest, by relating the number of sequence tags identified for each of the chromosomes of interest in step (b) to the length of each of the chromosomes of interest; (ii) calculating a sequence tag density ratio for each normalizing chromosome sequence by relating the

number of sequence tags identified for the sequence in step (b) to the length of each normalizing chromosome; and (iii) using the sequence tag density ratios calculated in steps (i) and (ii) to calculate a single chromosome dose for each of the chromosomes of interest, wherein the chromosome dose is calculated as the ratio of the sequence tag density ratio for each of the chromosomes of interest and the sequence tag density ratio for the normalizing chromosome sequence for each of the chromosomes of interest.

In another embodiment, a method is provided for determining the presence or absence of any four or more different complete fetal chromosomal aneuploidies in a maternal test sample comprising fetal and maternal nucleic acid molecules. The steps of the method comprise (a) obtaining sequence information for the fetal and maternal nucleic acids in the maternal test sample; (b) using the sequence information to identify a number of sequence tags for each of any four or more chromosomes of interest selected from chromosomes 1-22, X and Y and to identify a number of sequence tags for a normalizing chromosome sequence for each of the any four or more chromosomes of interest; (c) using the number of sequence tags identified for each of the any four or more chromosomes of interest and the number of sequence tags identified for each normalizing chromosome to calculate a single chromosome dose for each of the any four or more chromosomes of interest; and (d) comparing each of the single chromosome doses for each of the any four or more chromosomes of interest to a threshold value for each of the four or more chromosomes of interest, and thereby determining the presence or absence of any four or more complete different fetal chromosomal aneuploidies in the maternal test sample, wherein the any four or more chromosomes of interest selected from chromosomes 1-22, X, and Y comprise at least twenty chromosomes selected from chromosomes 1-22, X, and Y, and wherein the presence or absence of at least twenty different complete fetal chromosomal aneuploidies is determined. Step (a) can comprise sequencing at least a portion of the nucleic acid molecules of a test sample to obtain said sequence information for the fetal and maternal nucleic acid molecules of the test sample. In some embodiments, step (c) comprises calculating a single chromosome dose for each of the chromosomes of interest as the ratio of the number of sequence tags identified for each of the chromosomes of interest and the number of sequence tags identified for the normalizing chromosome sequence for each of the chromosomes of interest. In some other embodiments, step (c) comprises (i) calculating a sequence tag density ratio for each of the chromosomes of interest, by relating the number of sequence tags identified for each of the chromosomes of interest in step (b) to the length of each of the chromosomes of interest; (ii) calculating a sequence tag density ratio for each normalizing chromosome sequence by relating the number of sequence tags identified for the normalizing chromosome sequence in step (b) to the length of each normalizing chromosome; and (iii) using the sequence tag density ratios calculated in steps (i) and (ii) to calculate a single chromosome dose for each of the chromosomes of interest, wherein the chromosome dose is calculated as the ratio of the sequence tag density ratio for each of the

chromosomes of interest and the sequence tag density ratio for the normalizing chromosome sequence for each of the chromosomes of interest.

In another embodiment, a method is provided for determining the presence or absence of any four or more different complete fetal chromosomal aneuploidies in a maternal test sample comprising fetal and maternal nucleic acid molecules. The steps of the method comprise (a) obtaining sequence information for the fetal and maternal nucleic acids in the maternal test sample; (b) using the sequence information to identify a number of sequence tags for each of any four or more chromosomes of interest selected from chromosomes 1-22, X and Y and to identify a number of sequence tags for a normalizing chromosome sequence for each of the any four or more chromosomes of interest; (c) using the number of sequence tags identified for each of the any four or more chromosomes of interest and the number of sequence tags identified for each normalizing chromosome sequence to calculate a single chromosome dose for each of the any four or more chromosomes of interest; and (d) comparing each of the single chromosome doses for each of the any four or more chromosomes of interest to a threshold value for each of the four or more chromosomes of interest, and thereby determining the presence or absence of any four or more complete different fetal chromosomal aneuploidies in the maternal test sample, wherein the any four or more chromosomes of interest selected from chromosomes 1-22, X, and Y is all of chromosomes 1-22, X, and Y, and wherein the presence or absence of complete fetal chromosomal aneuploidies of all of chromosomes 1-22, X, and Y is determined. Step (a) can comprise sequencing at least a portion of the nucleic acid molecules of a test sample to obtain said sequence information for the fetal and maternal nucleic acid molecules of the test sample. In some embodiments, step (c) comprises calculating a single chromosome dose for each of the chromosomes of interest as the ratio of the number of sequence tags identified for each of the chromosomes of interest and the number of sequence tags identified for the normalizing chromosome sequence for each of the chromosomes of interest. In some other embodiments, step (c) comprises (i) calculating a sequence tag density ratio for each of the chromosomes of interest, by relating the number of sequence tags identified for each of the chromosomes of interest in step (b) to the length of each of the chromosomes of interest; (ii) calculating a sequence tag density ratio for each normalizing chromosome sequence by relating the number of sequence tags identified for the normalizing chromosome sequence in step (b) to the length of each normalizing chromosome; and (iii) using the sequence tag density ratios calculated in steps (i) and (ii) to calculate a single chromosome dose for each of the chromosomes of interest, wherein the chromosome dose is calculated as the ratio of the sequence tag density ratio for each of the chromosomes of interest and the sequence tag density ratio for the normalizing chromosome sequence for each of the chromosomes of interest.

In any of the embodiments above, the normalizing chromosome sequence may be a single chromosome selected from chromosomes 1-22, X, and Y. Alternatively, the normalizing chromosome sequence may be a group of chromosomes selected from chromosomes 1-22, X, and Y.

In another embodiment, a method is provided for determining the presence or absence of any one or more different complete fetal chromosomal aneuploidies in a maternal test sample comprising fetal and maternal nucleic acids. The steps of the method comprise: (a) obtaining sequence information for the fetal and maternal nucleic acids in the sample; (b) using the sequence information to identify a number of sequence tags for each of any one or more chromosomes of interest selected from chromosomes 1-22, X and Y and to identify a number of sequence tags for a normalizing segment sequence for each of any one or more chromosomes of interest; (c) using the number of sequence tags identified for each of any one or more chromosomes of interest and the number of sequence tags identified for the normalizing segment sequence to calculate a single chromosome dose for each of any one or more chromosomes of interest; and (d) comparing each of the single chromosome doses for each of any one or more chromosomes of interest to a threshold value for each of the one or more chromosomes of interest, and thereby determining the presence or absence of one or more different complete fetal chromosomal aneuploidies in the sample. Step (a) can comprise sequencing at least a portion of the nucleic acid molecules of a test sample to obtain said sequence information for the fetal and maternal nucleic acid molecules of the test sample.

In some embodiments, step (c) comprises calculating a single chromosome dose for each of the chromosomes of interest as the ratio of the number of sequence tags identified for each of the chromosomes of interest and the number of sequence tags identified for the normalizing segment sequence for each of the chromosomes of interest. In some other embodiments, step (c) comprises (i) calculating a sequence tag density ratio for each of chromosomes of interest, by relating the number of sequence tags identified for each chromosomes of interest in step (b) to the length of each of the chromosomes of interest; (ii) calculating a sequence tag density ratio for each normalizing segment sequence by relating the number of sequence tags identified for the normalizing segment sequence in step (b) to the length of each the normalizing chromosomes; and (iii) using the sequence tag density ratios calculated in steps (i) and (ii) to calculate a single chromosome dose for each of said chromosomes of interest, wherein said chromosome dose is calculated as the ratio of the sequence tag density ratio for each of the chromosomes of interest and the sequence tag density ratio for the normalizing segment sequence for each of the chromosomes of interest.

In another embodiment, a method is provided for determining the presence or absence of any one or more different complete fetal chromosomal aneuploidies in a maternal test sample comprising fetal and maternal nucleic acids. The steps of the method comprise: (a) obtaining sequence information for the fetal and maternal nucleic acids in the sample; (b) using the sequence information to identify a number of sequence tags for each of any one or more chromosomes of interest selected from chromosomes 1-22, X and Y and to identify a number of sequence tags for a normalizing segment sequence for each of any one or more chromosomes of interest; (c) using the number of sequence tags identified for each of any one or more chromosomes of interest and the number of

sequence tags identified for the normalizing segment sequence to calculate a single chromosome dose for each of any one or more chromosomes of interest; and (d) comparing each of the single chromosome doses for each of any one or more chromosomes of interest to a threshold value for each of the one or more chromosomes of interest, and thereby determining the presence or absence of one or more different complete fetal chromosomal aneuploidies in the sample, wherein the any one or more chromosomes of interest selected from chromosomes 1-22, X, and Y comprise at least twenty chromosomes selected from chromosomes 1-22, X, and Y, and wherein the presence or absence of at least twenty different complete fetal chromosomal aneuploidies is determined. Step (a) can comprise sequencing at least a portion of the nucleic acid molecules of a test sample to obtain said sequence information for the fetal and maternal nucleic acid molecules of the test sample. In some embodiments, step (c) comprises calculating a single chromosome dose for each of the chromosomes of interest as the ratio of the number of sequence tags identified for each of the chromosomes of interest and the number of sequence tags identified for the normalizing segment sequence for each of the chromosomes of interest. In some other embodiments, step (c) comprises (i) calculating a sequence tag density ratio for each of chromosomes of interest, by relating the number of sequence tags identified for each chromosomes of interest in step (b) to the length of each of the chromosomes of interest; (ii) calculating a sequence tag density ratio for each normalizing segment sequence by relating the number of sequence tags identified for the normalizing segment sequence in step (b) to the length of each the normalizing chromosomes; and (iii) using the sequence tag density ratios calculated in steps (i) and (ii) to calculate a single chromosome dose for each of said chromosomes of interest, wherein said chromosome dose is calculated as the ratio of the sequence tag density ratio for each of the chromosomes of interest and the sequence tag density ratio for the normalizing segment sequence for each of the chromosomes of interest.

In another embodiment, a method is provided for determining the presence or absence of any one or more different complete fetal chromosomal aneuploidies in a maternal test sample comprising fetal and maternal nucleic acids. The steps of the method comprise: (a) obtaining sequence information for the fetal and maternal nucleic acids in the sample; (b) using the sequence information to identify a number of sequence tags for each of any one or more chromosomes of interest selected from chromosomes 1-22, X and Y and to identify a number of sequence tags for a normalizing segment sequence for each of any one or more chromosomes of interest; (c) using the number of sequence tags identified for each of any one or more chromosomes of interest and the number of sequence tags identified for the normalizing segment sequence to calculate a single chromosome dose for each of any one or more chromosomes of interest; and (d) comparing each of the single chromosome doses for each of any one or more chromosomes of interest to a threshold value for each of the one or more chromosomes of interest, and thereby determining the presence or absence of one or more different complete fetal chromosomal aneuploidies in the sample, wherein the any one or

more chromosomes of interest selected from chromosomes 1-22, X, and Y is all of chromosomes 1-22, X, and Y, and wherein the presence or absence of complete fetal chromosomal aneuploidies of all of chromosomes 1-22, X, and Y is determined. Step (a) can comprise sequencing at least a portion of the nucleic acid molecules of a test sample to obtain said sequence information for the fetal and maternal nucleic acid molecules of the test sample. In some embodiments, step (c) comprises calculating a single chromosome dose for each of the chromosomes of interest as the ratio of the number of sequence tags identified for each of the chromosomes of interest and the number of sequence tags identified for the normalizing segment sequence for each of the chromosomes of interest. In some other embodiments, step (c) comprises (i) calculating a sequence tag density ratio for each of chromosomes of interest, by relating the number of sequence tags identified for each chromosomes of interest in step (b) to the length of each of the chromosomes of interest; (ii) calculating a sequence tag density ratio for each normalizing segment sequence by relating the number of sequence tags identified for the normalizing segment sequence in step (b) to the length of each the normalizing chromosomes; and (iii) using the sequence tag density ratios calculated in steps (i) and (ii) to calculate a single chromosome dose for each of said chromosomes of interest, wherein said chromosome dose is calculated as the ratio of the sequence tag density ratio for each of the chromosomes of interest and the sequence tag density ratio for the normalizing segment sequence for each of the chromosomes of interest.

In any one of the embodiments above, the different complete chromosomal aneuploidies are selected from complete chromosomal trisomies, complete chromosomal monosomies and complete chromosomal polysomies. The different complete chromosomal aneuploidies are selected from complete aneuploidies of any one of chromosome 1-22, X, and Y. For example, the said different complete fetal chromosomal aneuploidies are selected from trisomy 2, trisomy 8, trisomy 9, trisomy 21, trisomy 13, trisomy 16, trisomy 18, trisomy 22, 47,XXY, 47,XXX, 47,XYY, and monosomy X.

In any one of the embodiments above, steps (a)-(d) are repeated for test samples from different

maternal subjects, and the method comprises determining the presence or absence of any four or more different complete fetal chromosomal aneuploidies in each of the test samples.

In any one of the embodiments above, the method can further comprise calculating a normalized chromosome value (NCV), wherein the NCV relates the chromosome dose to the mean of the corresponding chromosome dose in a set of qualified samples as:

$$NCV_{ij} = \frac{x_{ij} - \hat{\mu}_j}{\hat{\sigma}_j}$$

where $\hat{\mu}_j$ and $\hat{\sigma}_j$ are the estimated mean and standard deviation, respectively, for the j-th chromosome dose in a set of qualified samples, and x_{ij} is the observed j-th chromosome dose for test sample i.

In another embodiment, a method is provided for determining the presence or absence of different partial fetal chromosomal aneuploidies in a maternal test sample comprising fetal and maternal nucleic acids. The steps of the method comprise: (a) obtaining sequence information for the fetal and maternal nucleic acids in the sample; (b) using the sequence information to identify a number of sequence tags for each of any one or more segments of any one or more chromosomes of interest selected from chromosomes 1-22, X, and Y and to identify a number of sequence tags for a normalizing segment sequence for each of any one or more segments of any one or more chromosomes of interest; (c) using the number of sequence tags identified for each of any one or more segments of any one or more chromosomes of interest and said number of sequence tags identified for the normalizing segment sequence to calculate a single segment dose for each of said any one or more segments of any one or more chromosomes of interest; and (d) comparing each of the single segment doses for each of any one or more segments of any one or more chromosomes of interest to a threshold value for each of any one or more chromosomal segments of any one or more chromosome of interest, and thereby determining the presence or absence of one or more different partial fetal chromosomal aneuploidies in the sample. Step (a) can comprise sequencing at least a portion of the nucleic acid molecules of a test sample to obtain said sequence information for the fetal and maternal nucleic acid molecules of the test sample.

In some embodiments, step (c) comprises calculating a single segment dose for each of any one or more segments of any one or more chromosomes of interest as the ratio of the number of sequence tags identified for each of any one or more segments of any one or more chromosomes of interest and the number of sequence tags identified for the normalizing segment sequence for each of the any one or more segments of any one or more chromosomes of interest. In some other embodiments, step (c) comprises (i) calculating a sequence tag density ratio for each of segment of interest, by relating the number of sequence tags identified for each segment of interest in step (b) to the length of each of the segment of interest; (ii) calculating a sequence tag density ratio for each normalizing segment sequence by relating the number of sequence tags identified for the normalizing segment sequence in step (b) to the length of each the normalizing segment sequence; and (iii) using the sequence tag density ratios calculated in steps (i) and (ii) to calculate a single segment dose for each segment of interest, wherein the segment dose is calculated as the ratio of the sequence tag density ratio for each of the segments of interest and the sequence tag density ratio for the normalizing segment sequence for each of the segments of interest. The method can further comprise calculating a normalized segment value (NSV), wherein the NSV relates said segment dose to the mean of the corresponding segment dose in a set of qualified samples as:

$$NSV_{ij} = \frac{x_{ij} - \hat{\mu}_j}{\hat{\sigma}_j}$$

where $\hat{\mu}_j$ and $\hat{\sigma}_j$ are the estimated mean and standard deviation, respectively, for the j-th segment dose in a set of qualified samples, and x_{ij} is the observed j-th segment dose for test sample i.

In embodiments of the method described whereby a chromosome dose or a segment dose is determined using a normalizing segment sequence, the normalizing segment sequence may be a single segment of any one or more of chromosomes 1-22, X, and Y. Alternatively, the normalizing segment sequence may be a group of segments of any one or more of chromosomes 1-22, X, and Y.

Steps (a)-(d) of the method for determining the presence or absence of a partial fetal chromosomal aneuploidy are repeated for test samples from different maternal subjects, and the method comprises determining the presence or absence of different partial fetal chromosomal aneuploidies in each of said samples. Partial fetal chromosomal aneuploidies that can be determined according to the method include partial aneuploidies of any segment of any chromosome. The partial aneuploidies can be selected from partial duplications, partial multiplications, partial insertions and partial deletions. Examples of partial aneuploidies that can be determined according to the method include partial monosomy of chromosome 1, partial monosomy of chromosome 4, partial monosomy of chromosome 5, partial monosomy of chromosome 7, partial monosomy of chromosome 11, partial monosomy of chromosome 15, partial monosomy of chromosome 17, partial monosomy of chromosome 18, and partial monosomy of chromosome 22.

In any one of the embodiments described above, the test sample may be a maternal sample selected from blood, plasma, serum, urine and saliva samples. In any one of the embodiments, the test sample is may be plasma sample. The nucleic acid molecules of the maternal sample are a mixture of fetal and maternal cell-free DNA molecules. Sequencing of the nucleic acids can be performed using next generation sequencing (NGS). In some embodiments, sequencing is massively parallel sequencing using sequencing-by-synthesis with reversible dye terminators. In other embodiments, sequencing is sequencing-by-ligation. In yet other embodiments, sequencing is single molecule sequencing. Optionally, an amplification step is performed prior to sequencing.

In another embodiment, a method is provided for determining the presence or absence of any twenty or more different complete fetal chromosomal aneuploidies in a maternal plasma test sample comprising a mixture of fetal and maternal cell-free DNA molecules. The steps of the method comprise: (a) sequencing at least a portion of the cell-free DNA molecules to obtain sequence information for the fetal and maternal cell-free DNA molecules in the sample; (b) using the sequence information to identify a number of sequence tags for each of any twenty or more chromosomes of interest selected from chromosomes 1-22, X, and Y and to identify a number of sequence tags for a normalizing chromosome for each of said twenty or more chromosomes of interest; (c) using the number of sequence tags identified for each of the twenty or more chromosomes of interest and the number of sequence tags identified for each normalizing chromosome to calculate a single chromosome dose for each of the twenty or more chromosomes of interest; and (d) comparing each of

the single chromosome doses for each of the twenty or more chromosomes of interest to a threshold value for each of the twenty or more chromosomes of interest, and thereby determining the presence or absence of any twenty or more different complete fetal chromosomal aneuploidies in the sample.

In another embodiment, the invention provides a method for identifying copy number variation (CNV) of a sequence of interest *e.g.* a clinically relevant sequence, in a test sample comprising the steps of: (a) obtaining a test sample and a plurality of qualified samples, said test sample comprising test nucleic acid molecules and said plurality of qualified samples comprising qualified nucleic acid molecules; (b) obtaining sequence information for said fetal and maternal nucleic acids in said sample; (c) based on said sequencing of said qualified nucleic acid molecules, calculating a qualified sequence dose for said qualified sequence of interest in each of said plurality of qualified samples, wherein said calculating a qualified sequence dose comprises determining a parameter for said qualified sequence of interest and at least one qualified normalizing sequence; (d) based on said qualified sequence dose, identifying at least one qualified normalizing sequence, wherein said at least one qualified normalizing sequence has the smallest variability and/or the greatest differentiability in sequence dose in said plurality of qualified samples; (e) based on said sequencing of said nucleic acid molecules in said test sample, calculating a test sequence dose for said test sequence of interest, wherein said calculating a test sequence dose comprises determining a parameter for said test sequence of interest and at least one normalizing test sequence, and wherein said at least one normalizing test sequence corresponds to said at least one qualified normalizing sequence; (f) comparing said test sequence dose to at least one threshold value; and (g) assessing said copy number variation of said sequence of interest in said test sample based on the outcome of step (f). In one embodiment, the parameter for said qualified sequence of interest and at least one qualified normalizing sequence relates the number of sequence tags mapped to said qualified sequence of interest to the number of tags mapped to said qualified normalizing sequence, and wherein said parameter for said test sequence of interest and at least one normalizing test sequence relates the number of sequence tags mapped to said test sequence of interest to the number of tags mapped to said normalizing test sequence. In some embodiments, step (b) comprises sequencing at least a portion of the qualified and test nucleic acid molecules, wherein sequencing comprises providing a plurality of mapped sequence tags for a test and a qualified sequence of interest, and for at least one test and at least one qualified normalizing sequence; sequencing at least a portion of said nucleic acid molecules of the test sample to obtain the sequence information for the fetal and maternal nucleic acid molecules of the test sample. In some embodiments, the sequencing step is performed using next generation sequencing method. In some embodiments, the sequencing method may be a massively parallel sequencing method that uses sequencing-by-synthesis with reversible dye terminators. In other embodiments, the sequencing method is sequencing-by-ligation. In some embodiments, sequencing comprises an amplification. In other embodiments, sequencing is single molecule

sequencing. The CNV of a sequence of interest is an aneuploidy, which can be a chromosomal or a partial aneuploidy. In some embodiments, the chromosomal aneuploidy is selected from trisomy 2, trisomy 8, trisomy 9, trisomy 16, trisomy 21, trisomy 13, trisomy 18, trisomy 22, 47,XXY, 47,XXX, 47,XYY, and monosomy X. In other embodiments, the partial aneuploidy is a partial chromosomal deletion or a partial chromosomal insertion. In some embodiments, the CNV identified by the method is a chromosomal or partial aneuploidy associated with cancer. In some embodiments, the test and qualified sample are biological fluid samples *e.g.* plasma samples, obtained from a pregnant subject such as a pregnant human subject. In other embodiments, a test and qualified biological fluid samples *e.g.* plasma samples, are obtained from a subject that is known or is suspected of having cancer.

Although the examples herein concern humans and the language is primarily directed to human concerns, the concept of this invention is applicable to genomes from any plant or animal.

INCORPORATION BY REFERENCE

All patents, patent applications, and other publications, including all sequences disclosed within these references, referred to herein are expressly incorporated by reference, to the same extent as if each individual publication, patent or patent application was specifically and individually indicated to be incorporated by reference. All documents cited are, in relevant part, incorporated herein by reference. However, the citation of any document is not to be construed as an admission that it is prior art with respect to the present invention.

BRIEF DESCRIPTION OF THE DRAWINGS

The novel features of the invention are set forth with particularity in the appended claims. A better understanding of the features and advantages of the present invention will be obtained by reference to the following detailed description that sets forth illustrative embodiments, in which the principles of the invention are utilized, and the accompanying drawings of which:

Figure 1 is a flowchart of a method **100** for determining the presence or absence of a copy number variation in a test sample comprising a mixture of nucleic acids.

Figure 2 illustrates the distribution of the chromosome dose for chromosome 21 determined from sequencing cfDNA extracted from a set of 48 blood samples obtained from human subjects each pregnant with a male or a female fetus. Chromosome 21 doses for qualified *i.e.* normal for chromosome 21 (O), and trisomy 21 test samples are shown (**Δ**) for chromosomes 1-12 and X (Figure 2A), and for chromosomes 1-22 and X (Figure 2B).

Figure 3 illustrates the distribution of the chromosome dose for chromosome 18 determined from sequencing cfDNA extracted from a set of 48 blood samples obtained from human subjects each pregnant with a male or a female fetus. Chromosome 18 doses for qualified *i.e.* normal for

chromosome 18 (O), and trisomy 18 (Δ) test samples are shown for chromosomes 1-12 and X (Figure 3A), and for chromosomes 1-22 and X (Figure 3B).

Figure 4 illustrates the distribution of the chromosome dose for chromosome 13 determined from sequencing cfDNA extracted from a set of 48 blood samples obtained from human subjects each pregnant with a male or a female fetus. Chromosome 13 doses for qualified *i.e.* normal for chromosome 13 (O), and trisomy 13 (Δ) test samples are shown for chromosomes 1-12 and X (Figure 4A), and for chromosomes 1-22 and X (Figure 4B).

Figure 5 illustrates the distribution of the chromosome doses for chromosome X determined from sequencing cfDNA extracted from a set of 48 test blood samples obtained from human subjects each pregnant with a male or a female fetus. Chromosome X doses for males (46,XY; (O)), females (46,XX; (Δ)); monosomy X (45,X; (+)), and complex karyotypes (Cplx (X)) samples are shown for chromosomes 1-12 and X (Figure 5A), and for chromosomes 1-22 and X (Figure 5B).

Figure 6 illustrates the distribution of the chromosome doses for chromosome Y determined from sequencing cfDNA extracted from a set of 48 test blood samples obtained from human subjects each pregnant with a male or a female fetus. Chromosome Y doses for males (46,XY; (Δ)), females (46,XX; (O)); monosomy X (45,X; (+)), and complex karyotypes (Cplx (X)) samples are shown for chromosomes 1-12 (Figure 6A), and for chromosomes 1-22 (Figure 6B).

Figure 7 shows the coefficient of variation (CV) for chromosomes 21 (■), 18 (●) and 13 (▲) that was determined from the doses shown in Figures 2, 3, and 4, respectively.

Figure 8 shows the coefficient of variation (CV) for chromosomes X (■) and Y (●) that was determined from the doses shown in Figures 5 and 6, respectively.

Figure 9 shows the cumulative distribution of GC fraction by human chromosome. The vertical axis represents the frequency of the chromosome with GC content below the value shown on the horizontal axis.

Figure 10 illustrates the sequence doses (Y-axis) for a segment of chromosome 11 (81000082-103000103bp) determined from sequencing cfDNA extracted from a set of 7 qualified samples (O) obtained and 1 test sample (♦) from pregnant human subjects. A sample from a subject carrying a fetus with a partial aneuploidy of chromosome 11 (♦) was identified.

Figure 11 illustrates the distribution of normalized chromosome doses for chromosome 21 (A), chromosome 18 (B), chromosome 13 (C), chromosome X (D) and chromosome Y (E) relative to the standard deviation of the mean (Y-axis) for the corresponding chromosomes in the unaffected samples.

Figure 12 shows normalized chromosome values for chromosomes 21 (O), 18 (Δ), and 13 (□) determined in samples from training set 1 using normalizing chromosomes as described in Example 6.

Figure 13 shows normalized chromosome values for chromosomes 21 (O), 18 (Δ), and 13 (□) determined in samples from test set 1 using normalizing chromosomes as described in Example 6.

Figure 14 shows normalized chromosome values for chromosomes 21 (O) and 18 (Δ) determined in samples from test set 1 using the normalizing method of Chiu *et al.* (normalizes the number of sequence tags identified for the chromosome of interest with the number of sequence tags obtained for the remaining chromosomes in the sample; see elsewhere herein Example 7).

Figure 15 shows normalized chromosome values for chromosomes 21 (O), 18 (Δ), and 13 (□) determined in samples from training set 1 using systematically determined normalizing chromosomes (as described in Example 7).

Figure 16 shows normalized chromosome values for chromosomes 21 (O), 18 (Δ), and 13 (□) determined in samples from test set 1 using systematically determined normalizing chromosomes (as described in Example 7).

Figure 17 shows normalized chromosome values for chromosome 9 (O) determined in samples from test set 1 using systematically determined normalizing chromosomes (as described in Example 7).

Figure 18 shows normalized chromosome values for chromosomes X (X-axis) and Y (Y-axis). The arrows point to the 5 (Figure 18A) and 3 (Figure 18B) monosomy X samples that were identified in the training and test sets, respectively, as described in Example 7.

Figure 19 shows normalized chromosome values for chromosomes 1-22 determined in samples from test set 1 using systematically determined normalizing chromosomes (as described in Example 7).

DETAILED DESCRIPTION OF THE INVENTION

The invention provides a method for determining copy number variations (CNV) of a sequence of interest in a test sample that comprises a mixture of nucleic acids that are known or are suspected to differ in the amount of one or more sequence of interest. Sequences of interest include genomic sequences ranging from kilobases (kb) to megabases (Mb) to entire chromosomes that are known or are suspected to be associated with a genetic or a disease condition. Examples of sequences of interest include chromosomes associated with well known aneuploidies *e.g.* trisomy 21, and segments of chromosomes that are multiplied in diseases such as cancer *e.g.* partial trisomy 8 in acute myeloid leukemia. CNV that can be determined according to the present method include monosomies and trisomies of any one or more of autosomes 1-22, and of sex chromosomes X and Y *e.g.* 45,X, 47,XXX, 47,XXY and 47,XYY, other chromosomal polysomies *i.e.* tetrasomy and pentasomies including but not limited to XXXX, XXXXX, XXXXY and XYYYY, and deletions and/or duplications of segments of any one or more of the chromosomes.

The method comprises a statistical approach that accounts for accrued variability stemming from process-related, interchromosomal (intra-run), and inter-sequencing (inter-run) variability. The

method is applicable to determining CNV of any fetal aneuploidy, and CNVs known or suspected to be associated with a variety of medical conditions.

Unless otherwise indicated, the practice of the present invention involves conventional techniques commonly used in molecular biology, microbiology, protein purification, protein engineering, protein and DNA sequencing, and recombinant DNA fields, which are within the skill of the art. Such techniques are known to those of skill in the art and are described in numerous texts and reference works (See *e.g.*, Sambrook et al., "Molecular Cloning: A Laboratory Manual", Third Edition (Cold Spring Harbor), [2001]); and Ausubel et al., "Current Protocols in Molecular Biology" [1987]).

Numeric ranges are inclusive of the numbers defining the range. It is intended that every maximum numerical limitation given throughout this specification includes every lower numerical limitation, as if such lower numerical limitations were expressly written herein. Every minimum numerical limitation given throughout this specification will include every higher numerical limitation, as if such higher numerical limitations were expressly written herein. Every numerical range given throughout this specification will include every narrower numerical range that falls within such broader numerical range, as if such narrower numerical ranges were all expressly written herein.

The headings provided herein are not limitations of the various aspects or embodiments of the invention which can be had by reference to the Specification as a whole. Accordingly, as indicated above, the terms defined immediately below are more fully defined by reference to the specification as a whole.

Unless defined otherwise herein, all technical and scientific terms used herein have the same meaning as commonly understood by one of ordinary skill in the art to which this invention belongs. Various scientific dictionaries that include the terms included herein are well known and available to those in the art. Although any methods and materials similar or equivalent to those described herein find use in the practice or testing of the present invention, some preferred methods and materials are described. Accordingly, the terms defined immediately below are more fully described by reference to the Specification as a whole. It is to be understood that this invention is not limited to the particular methodology, protocols, and reagents described, as these may vary, depending upon the context they are used by those of skill in the art.

DEFINITIONS

As used herein, the singular terms "a", "an," and "the" include the plural reference unless the context clearly indicates otherwise. Unless otherwise indicated, nucleic acids are written left to right in 5' to 3' orientation and amino acid sequences are written left to right in amino to carboxy orientation, respectively.

The term “assessing” herein refers to characterizing the status of a chromosomal aneuploidy by one of three types of calls: “normal”, “affected”, and “no-call”. For example, in the presence of

trisomy the "normal" call is determined by the value of a parameter *e.g.* a test chromosome dose that is below a user-defined threshold of reliability, the "affected" call is determined by a parameter *e.g.* a test chromosome dose, that is above a user-defined threshold of reliability, and the "no-call" result is determined by a parameter *e.g.* a test chromosome dose, that lies between the user-defined thresholds of reliability for making a "normal" or an "affected" call.

The term "copy number variation" herein refers to variation in the number of copies of a nucleic acid sequence that is 1 kb or larger present in a test sample in comparison with the copy number of the nucleic acid sequence present in a qualified sample. A "copy number variant" refers to the 1 kb or larger sequence of nucleic acid in which copy-number differences are found by comparison of a sequence of interest in test sample with that present in a qualified sample. Copy number variants/variations include deletions, including microdeletions, insertions, including microinsertions, duplications, multiplications, inversions, translocations and complex multi-site variants. CNV encompass chromosomal aneuploidies and partial aneuploidies.

The term "aneuploidy" herein refers to an imbalance of genetic material caused by a loss or gain of a whole chromosome, or part of a chromosome.

The terms "chromosomal aneuploidy" and "complete chromosomal aneuploidy" herein refer to an imbalance of genetic material caused by a loss or gain of a whole chromosome, and includes germline aneuploidy and mosaic aneuploidy.

The terms "partial aneuploidy" and "partial chromosomal aneuploidy" herein refer to an imbalance of genetic material caused by a loss or gain of part of a chromosome *e.g.* partial monosomy and partial trisomy, and encompasses imbalances resulting from translocations, deletions and insertions.

The term "aneuploid sample" herein refers to a sample indicative of a subject whose chromosomal content is not euploid, *i.e.* the sample is indicative of a subject with an abnormal copy number of chromosomes.

The term "aneuploid chromosome" herein refers to a chromosome that is known or determined to be present in a sample in an abnormal copy number.

The term "plurality" is used herein in reference to a number of nucleic acid molecules or sequence tags that is sufficient to identify significant differences in copy number variations (*e.g.* chromosome doses) in test samples and qualified samples using in the methods of the invention. In some embodiments, at least about 3×10^6 sequence tags, at least about 5×10^6 sequence tags, at least about 8×10^6 sequence tags, at least about 10×10^6 sequence tags, at least about 15×10^6 sequence tags, at least about 20×10^6 sequence tags, at least about 30×10^6 sequence tags, at least about 40×10^6 sequence tags, or at least about 50×10^6 sequence tags comprising between 20 and 40bp reads are obtained for each test sample.

The terms "polynucleotide", "nucleic acid" and “nucleic acid molecules” are used interchangeably and refer to a covalently linked sequence of nucleotides (i.e., ribonucleotides for RNA and deoxyribonucleotides for DNA) in which the 3' position of the pentose of one nucleotide is joined by a phosphodiester group to the 5' position of the pentose of the next, include sequences of any form of nucleic acid, including, but not limited to RNA, DNA and cfDNA molecules. The term "polynucleotide" includes, without limitation, single- and double-stranded polynucleotide.

The term “portion” is used herein in reference to the amount of sequence information of fetal and maternal nucleic acid molecules in a biological sample that in sum amount to less than the sequence information of <1 human genome.

The term “test sample” herein refers to a sample comprising a mixture of nucleic acids comprising at least one nucleic acid sequence whose copy number is suspected of having undergone variation. Nucleic acids present in a test sample are referred to as “test nucleic acids”.

The term “qualified sample” herein refers to a sample comprising a mixture of nucleic acids that are present in a known copy number to which the nucleic acids in a test sample are compared, and it is a sample that is normal *i.e.* not aneuploid, for the sequence of interest *e.g.* a qualified sample used for identifying a normalizing chromosome for chromosome 21 is a sample that is not a trisomy 21 sample.

The term “training set” herein refers to a set of samples that can comprise affected and unaffected samples. The unaffected samples in a training set are used as the qualified samples to identify normalizing sequences, *e.g.* normalizing chromosomes, and the chromosome doses of unaffected samples are used to set the thresholds for each of the sequences , *e.g.* chromosomes, of interest. The affected samples in a training set can be used to verify that affected test samples can be easily differentiated from unaffected samples.

The term “qualified nucleic acid” is used interchangeably with “qualified sequence” is a sequence against which the amount of a test sequence or test nucleic acid is compared. A qualified sequence is one present in a biological sample preferably at a known representation *i.e.* the amount of a qualified sequence is known. A “qualified sequence of interest” is a qualified sequence for which the amount is known in a qualified sample, and is a sequence that is associated with a difference in sequence representation in an individual with a medical condition.

The term “sequence of interest” herein refers to a nucleic acid sequence that is associated with a difference in sequence representation in healthy versus diseased individuals. A sequence of interest can be a sequence on a chromosome that is misrepresented *i.e.* over- or under-represented, in a disease or genetic condition. A sequence of interest may also be a portion of a chromosome *i.e.* chromosome segment, or a chromosome. For example, a sequence of interest can be a chromosome that is over-represented in an aneuploidy condition, or a gene encoding a tumor-suppressor that is under-represented in a cancer. Sequences of interest include sequences that are over- or under- represented

in the total population, or a subpopulation of cells of a subject. A "qualified sequence of interest" is a sequence of interest in a qualified sample. A "test sequence of interest" is a sequence of interest in a test sample.

The term "normalizing sequence" herein refers to a sequence that displays a variability in the number of sequence tags that are mapped to it among samples and sequencing runs that best approximates that of the sequence of interest for which it is used as a normalizing parameter, and that can best differentiate an affected sample from one or more unaffected samples. A "normalizing chromosome" or "normalizing chromosome sequence" is an example of a "normalizing sequence". A "normalizing chromosome sequence" can be composed of a single chromosome or of a group of chromosomes. A "normalizing segment" is another example of a "normalizing sequence". A "normalizing segment sequence" can be composed of a single segment of a chromosome or it can be composed of two or more segments of the same or of different chromosomes.

The term "differentiability" herein refers to the characteristic of a normalizing chromosome that enables to distinguish one or more unaffected *i.e.* normal, samples from one or more affected *i.e.* aneuploid, samples.

The term "sequence dose" herein refers to a parameter that relates the sequence tag density of a sequence of interest to the tag density of a normalizing sequence. A "test sequence dose" is a parameter that relates the sequence tag density of a sequence of interest *e.g.* chromosome 21, to that of a normalizing sequence *e.g.* chromosome 9, determined in a test sample. Similarly, a "qualified sequence dose" is a parameter that relates the sequence tag density of a sequence of interest to that of a normalizing sequence determined in a qualified sample.

The term "sequence tag density" herein refers to the number of sequence reads that are mapped to a reference genome sequence *e.g.* the sequence tag density for chromosome 21 is the number of sequence reads generated by the sequencing method that are mapped to chromosome 21 of the reference genome. The term "sequence tag density ratio" herein refers to the ratio of the number of sequence tags that are mapped to a chromosome of the reference genome *e.g.* chromosome 21, to the length of the reference genome chromosome 21.

The term "Next Generation Sequencing (NGS)" herein refers to sequencing methods that allow for massively parallel sequencing of clonally amplified and of single nucleic acid molecules. Non-limiting examples of NGS include sequencing-by-synthesis using reversible dye terminators, and sequencing-by-ligation.

The term "parameter" herein refers to 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 the number of sequence tags mapped to a chromosome and the length of the chromosome to which the tags are mapped, is a parameter.

The terms "threshold value" and "qualified threshold value" herein refer to any number that is calculated using a qualifying data set and serves as a limit of diagnosis of a copy number variation *e.g.* an aneuploidy, in an organism. If a threshold is exceeded by results obtained from practicing the invention, a subject can be diagnosed with a copy number variation *e.g.* trisomy 21. Appropriate threshold values for the methods described herein can be identified by analyzing normalizing values (*e.g.* chromosome doses, NCVs or NSVs) calculated for a training set of samples. Threshold values can be identified using qualified (*i.e.* unaffected) samples in a training set which comprises both qualified (*i.e.* unaffected) samples and affected samples. The samples in the training set known to have chromosomal aneuploidies (*i.e.* the affected samples) can be used to confirm that the chosen thresholds are useful in differentiating affected from unaffected samples in a test set (see the Examples herein). The choice of a threshold is dependent on the level of confidence that the user wishes to have to make the classification. In some embodiments, the training set used to identify appropriate threshold values comprises at least 10, at least 20, at least 30, at least 40, at least 50, at least 60, at least 70, at least 80, at least 90, at least 100, at least 200, at least 300, at least 400, at least 500, at least 600, at least 700, at least 800, at least 900, at least 1000, at least 2000 , at least 3000 , at least 4000, or more qualified samples. It may advantageous to use larger sets of qualified samples to improve the diagnostic utility of the threshold values.

The term "normalizing value" herein refers to a numerical value that relates the number of sequence tags identified for the sequence (*e.g.* chromosome or chromosome segment) of interest to the number of sequence tags identified for the normalizing sequence (*e.g.* normalizing chromosome or normalizing chromosome segment). For example, a "normalizing value" can be a chromosome dose as described elsewhere herein, or it can be an NCV (Normalized Chromosome Value) as described elsewhere herein, or it can be an NSV (Normalized Segment Value) as described elsewhere herein.

The term "read" refers to a DNA sequence of sufficient length (*e.g.*, at least about 30 bp) that can be used to identify a larger sequence or region, e.g. that can be aligned and specifically assigned to a chromosome or genomic region or gene.

The term "sequence tag" is herein used interchangeably with the term "mapped sequence tag" to refer to a sequence read that has been specifically assigned *i.e.* mapped, to a larger sequence *e.g.* a reference genome, by alignment. Mapped sequence tags are uniquely mapped to a reference genome *i.e.* they are assigned to a single location to the reference genome. Tags that can be mapped to more than one location on a reference genome *i.e.* tags that do not map uniquely, are not included in the analysis.

As used herein, the terms "aligned", "alignment", or "aligning" refer to one or more sequences that are identified as a match in terms of the order of their nucleic acid molecules to a known sequence from a reference genome. Such alignment can be done manually or by a computer algorithm, examples including the Efficient Local Alignment of Nucleotide Data (ELAND) computer

program distributed as part of the Illumina Genomics Analysis pipeline. The matching of a sequence read in aligning can be a 100% sequence match or less than 100% (non-perfect match).

As used herein, the term “reference genome” refers to any particular known genome sequence, whether partial or complete, of any organism or virus which may be used to reference identified sequences from a subject. For example, a reference genome used for human subjects as well as many other organisms is found at the National Center for Biotechnology Information at www.ncbi.nlm.nih.gov. A “genome” refers to the complete genetic information of an organism or virus, expressed in nucleic acid sequences.

The term “clinically-relevant sequence” herein refers to a nucleic acid sequence that is known or is suspected to be associated or implicated with a genetic or disease condition. Determining the absence or presence of a clinically-relevant sequence can be useful in determining a diagnosis or confirming a diagnosis of a medical condition, or providing a prognosis for the development of a disease.

The term “derived” when used in the context of a nucleic acid or a mixture of nucleic acids, herein refers to the means whereby the nucleic acid(s) are obtained from the source from which they originate. For example, in one embodiment, a mixture of nucleic acids that is derived from two different genomes means that the nucleic acids *e.g.* cfDNA, were naturally released by cells through naturally occurring processes such as necrosis or apoptosis. In another embodiment, a mixture of nucleic acids that is derived from two different genomes means that the nucleic acids were extracted from two different types of cells from a subject.

The term “mixed sample” herein refers to a sample containing a mixture of nucleic acids, which are derived from different genomes.

The term “maternal sample” herein refers to a biological sample obtained from a pregnant subject *e.g.* a woman.

The term "biological fluid" herein refers to a liquid taken from a biological source and includes, for example, blood, serum, plasma, sputum, lavage fluid, cerebrospinal fluid, urine, semen, sweat, tears, saliva, and the like. As used herein, the terms "blood," "plasma" and "serum" expressly encompass fractions or processed portions thereof. Similarly, where a sample is taken from a biopsy, swab, smear, *etc.*, the "sample" expressly encompasses a processed fraction or portion derived from the biopsy, swab, smear, *etc.*

The terms “maternal nucleic acids” and “fetal nucleic acids” herein refer to the nucleic acids of a pregnant female subject and the nucleic acids of the fetus being carried by the pregnant female, respectively.

As used herein, the term “corresponding to” refers to a nucleic acid sequence *e.g.* a gene or a chromosome, that is present in the genome of different subjects, and which does not necessarily have

the same sequence in all genomes, but serves to provide the identity rather than the genetic information of a sequence of interest *e.g.* a gene or chromosome.

As used herein, the term "substantially cell free" encompasses preparations of the desired sample from which components that are normally associated with it are removed. For example, a plasma sample is rendered essentially cell free by removing blood cells *e.g.* red cells, which are normally associated with it. In some embodiments, substantially free samples are processed to remove cells that would otherwise contribute to the desired genetic material that is to be tested for a CNV.

As used herein, the term “fetal fraction” refers to the fraction of fetal nucleic acids present in a sample comprising fetal and maternal nucleic acid.

As used herein the term "chromosome" refers to the heredity-bearing gene carrier of a living cell which is derived from chromatin and which comprises DNA and protein components (especially histones). The conventional internationally recognized individual human genome chromosome numbering system is employed herein.

As used herein, the term “polynucleotide length” refers to the absolute number of nucleic acid molecules (nucleotides) in a sequence or in a region of a reference genome. The term “chromosome length” refers to the known length of the chromosome given in base pairs *e.g.* provided in the NCBI36/hg18 assembly of the human chromosome found on the world wide web at genome.ucsc.edu/cgi-bin/hgTracks?hgsid=167155613&chromInfoPage=

The term “subject” herein refers to a human subject as well as a non-human subject such as a mammal, an invertebrate, a vertebrate, a fungus, a yeast, a bacteria, and a virus. Although the examples herein concern humans and the language is primarily directed to human concerns, the concept of this invention is applicable to genomes from any plant or animal, and is useful in the fields of veterinary medicine, animal sciences, research laboratories and such.

The term “condition” herein refers to “medical condition” as a broad term that includes all diseases and disorders, but can include [injuries] and normal health situations, such as pregnancy, that might affect a person's health, benefit from medical assistance, or have implications for medical treatments.

The term “complete” is used herein in reference to a chromosomal aneuploidy to refer to a gain or loss of an entire chromosome.

The term “partial” when used in reference to a chromosomal aneuploidy herein refers to a gain or loss of a portion of a chromosome.

The term “mosaic” herein refers to denote the presence of two populations of cells with different karyotypes in one individual who has developed from a single fertilized egg. Mosaicism may result from a mutation during development which is propagated to only a subset of the adult cells.

The term “non-mosaic” herein refers to an organism *e.g.* a human fetus, composed of cell of one karyotypes.

The term "using a chromosome" when used in reference to determining a chromosome dose, herein refers to using the sequence information obtained for a chromosome *i.e.* the number of sequence tags obtained for a chromosome.

The term "sensitivity" is used herein is equal to the number of true positives divided by the sum of true positives and false negatives.

The term "specificity" is used herein is equal to the number of true negatives divided by the sum of true negatives and false positives.

The term "patient sample" herein refers to a biological sample obtained from a patient *i.e.* a recipient of medical attention, care or treatment. The patient sample can be any of the samples described herein. Preferably, the patient sample is obtained by non-invasive procedures *e.g.* peripheral blood sample or a stool sample.

The term "hypodiploid" herein refers to a chromosome number that is one or more lower than the normal haploid number of chromosomes characteristic for the species.

DESCRIPTION

The invention provides a method for determining copy number variations (CNV) of different sequences of interest in a test sample that comprises a mixture of nucleic acids derived from two different genomes, and which are known or are suspected to differ in the amount of one or more sequence of interest. Copy number variations determined by the method of the invention include gains or losses of entire chromosomes, alterations involving very large chromosomal segments that are microscopically visible, and an abundance of sub-microscopic copy number variation of DNA segments ranging from kilobases (kb) to megabases (Mb) in size. The method comprises a statistical approach that accounts for accrued variability stemming from process-related, interchromosomal and inter-sequencing variability. The method is applicable to determining CNV of any fetal aneuploidy, and CNVs known or suspected to be associated with a variety of medical conditions. CNV that can be determined according to the present method include trisomies and monosomies of any one or more of chromosomes 1-22, X and Y, other chromosomal polysomies, and deletions and/or duplications of segments of any one or more of the chromosomes, which can be detected by sequencing only once the nucleic acids of a test sample. Any aneuploidy can be determined from sequencing information that is obtained by sequencing only once the nucleic acids of a test sample.

CNV in the human genome significantly influence human diversity and predisposition to disease (Redon *et al.,* Nature 23:444-454 [2006], Shaikh *et al.* Genome Res 19:1682-1690 [2009]) . CNVs have been known to contribute to genetic disease through different mechanisms, resulting in either imbalance of gene dosage or gene disruption in most cases. In addition to their direct correlation with genetic disorders, CNVs are known to mediate phenotypic changes that can be deleterious. Recently, several studies have reported an increased burden of rare or de novo CNVs in

complex disorders such as Autism, ADHD, and schizophrenia as compared to normal controls, highlighting the potential pathogenicity of rare or unique CNVs (Sebat *et al.,* 316:445 - 449 [2007]; Walsh *et al.,* Science 320:539 – 543 [2008]). CNV arise from genomic rearrangements, primarily owing to deletion, duplication, insertion, and unbalanced translocation events.

The method described herein employs next generation sequencing technology (NGS) in which clonally amplified DNA templates or single DNA molecules are sequenced in a massively parallel fashion within a flow cell (*e.g.* as described in Volkerding *et al.* Clin Chem 55:641-658 [2009]; Metzker M Nature Rev 11:31-46 [2010]). In addition to high-throughput sequence information, NGS provides quantitative information, in that each sequence read is a countable "sequence tag" representing an individual clonal DNA template or a single DNA molecule. The sequencing technologies of NGS include pyrosequencing, sequencing-by-synthesis with reversible dye terminators, sequencing by oligonucleotide probe ligation and ion semiconductor sequencing. DNA from individual samples can be sequenced individually (*i.e.* singleplex sequencing) or DNA from multiple samples can be pooled and sequenced as indexed genomic molecules (*i.e.* multiplex sequencing) on a single sequencing run, to generate up to several hundred million reads of DNA sequences. Examples of sequencing technologies that can be used to obtain the sequence information according to the present method are described below.

Sequencing Methods

Some of the sequencing technologies are available commercially, such as the sequencing-by-hybridization platform from Affymetrix Inc. (Sunnyvale, CA) and the sequencing-by-synthesis platforms from 454 Life Sciences (Bradford, CT), Illumina/Solexa (Hayward, CA) and Helicos Biosciences (Cambridge, MA), and the sequencing-by-ligation platform from Applied Biosystems (Foster City, CA), as described below. In addition to the single molecule sequencing performed using sequencing-by-synthesis of Helicos Biosciences, other single molecule sequencing technologies include the SMRT™ technology of Pacific Biosciences, the Ion Torrent™ technology, and nanopore sequencing being developed for example, by Oxford Nanopore Technologies. While the automated Sanger method is considered as a 'first generation' technology, Sanger sequencing including the automated Sanger sequencing, can also be employed by the method of the invention. Additional sequencing methods nucleic acid imaging technologies *e.g.* atomic force microscopy (AFM) or transmission electron microscopy (TEM). Exemplary sequencing technologies are described below.

In one embodiment, the present method comprises obtaining sequence information for the nucleic acids in a test sample *e.g.* cfDNA in a maternal sample, using single molecule sequencing technology of the Helicos True Single Molecule Sequencing (tSMS) technology (*e.g.* as described in Harris T.D. et al., Science 320:106-109 [2008]). In the tSMS technique, a DNA sample is cleaved into strands of approximately 100 to 200 nucleotides, and a polyA sequence is added to the 3' end of each DNA strand. Each strand is labeled by the addition of a fluorescently labeled adenosine

nucleotide. The DNA strands are then hybridized to a flow cell, which contains millions of oligo-T capture sites that are immobilized to the flow cell surface. The templates can be at a density of about 100 million templates/cm^2. The flow cell is then loaded into an instrument, e.g., HeliScope™ sequencer, and a laser illuminates the surface of the flow cell, revealing the position of each template. A CCD camera can map the position of the templates on the flow cell surface. The template fluorescent label is then cleaved and washed away. The sequencing reaction begins by introducing a DNA polymerase and a fluorescently labeled nucleotide. The oligo-T nucleic acid serves as a primer. The polymerase incorporates the labeled nucleotides to the primer in a template directed manner. The polymerase and unincorporated nucleotides are removed. The templates that have directed incorporation of the fluorescently labeled nucleotide are discerned by imaging the flow cell surface. After imaging, a cleavage step removes the fluorescent label, and the process is repeated with other fluorescently labeled nucleotides until the desired read length is achieved. Sequence information is collected with each nucleotide addition step. Whole genome sequencing by single molecule sequencing technologies excludes PCR-based amplification in the preparation of the sequencing libraries, and the directness of sample preparation allows for direct measurement of the sample, rather than measurement of copies of that sample.

In another embodiment, the present method comprises obtaining sequence information for the nucleic acids in the test sample *e.g.* cfDNA in a maternal test sample, using the 454 sequencing (Roche) (*e.g.* as described in Margulies, M. et al. Nature 437:376-380 [2005]). 454 sequencing involves two steps. In the first step, DNA is sheared into fragments of approximately 300-800 base pairs, and the fragments are blunt-ended. Oligonucleotide adaptors are then ligated to the ends of the fragments. The adaptors serve as primers for amplification and sequencing of the fragments. The fragments can be attached to DNA capture beads, e.g., streptavidin-coated beads using, e.g., Adaptor B, which contains 5'-biotin tag. The fragments attached to the beads are PCR amplified within droplets of an oil-water emulsion. The result is multiple copies of clonally amplified DNA fragments on each bead. In the second step, the beads are captured in wells (pico-liter sized). Pyrosequencing is performed on each DNA fragment in parallel. Addition of one or more nucleotides generates a light signal that is recorded by a CCD camera in a sequencing instrument. The signal strength is proportional to the number of nucleotides incorporated. Pyrosequencing makes use of pyrophosphate (PPi) which is released upon nucleotide addition. PPi is converted to ATP by ATP sulfurylase in the presence of adenosine 5' phosphosulfate. Luciferase uses ATP to convert luciferin to oxyluciferin, and this reaction generates light that is measured and analyzed.

In another embodiment, the present method comprises obtaining sequence information for the nucleic acids in the test sample *e.g.* cfDNA in a maternal test sample, using the SOLiD™ technology (Applied Biosystems). In SOLiD™ sequencing-by-ligation, genomic DNA is sheared into fragments, and adaptors are attached to the 5' and 3' ends of the fragments to generate a fragment library.

Alternatively, internal adaptors can be introduced by ligating adaptors to the 5' and 3' ends of the fragments, circularizing the fragments, digesting the circularized fragment to generate an internal adaptor, and attaching adaptors to the 5' and 3' ends of the resulting fragments to generate a mate-paired library. Next, clonal bead populations are prepared in microreactors containing beads, primers, template, and PCR components. Following PCR, the templates are denatured and beads are enriched to separate the beads with extended templates. Templates on the selected beads are subjected to a 3' modification that permits bonding to a glass slide. The sequence can be determined by sequential hybridization and ligation of partially random oligonucleotides with a central determined base (or pair of bases) that is identified by a specific fluorophore. After a color is recorded, the ligated oligonucleotide is cleaved and removed and the process is then repeated.

In another embodiment, the present method comprises obtaining sequence information for the nucleic acids in the test sample *e.g.* cfDNA in a maternal test sample, using the single molecule, real-time (SMRT™) sequencing technology of Pacific Biosciences. In SMRT sequencing, the continuous incorporation of dye-labeled nucleotides is imaged during DNA synthesis. Single DNA polymerase molecules are attached to the bottom surface of individual zero-mode wavelength detectors (ZMW detectors) that obtain sequence information while phospholinked nucleotides are being incorporated into the growing primer strand. A ZMW is a confinement structure which enables observation of incorporation of a single nucleotide by DNA polymerase against the background of fluorescent nucleotides that rapidly diffuse in an out of the ZMW (in microseconds). It takes several milliseconds to incorporate a nucleotide into a growing strand. During this time, the fluorescent label is excited and produces a fluorescent signal, and the fluorescent tag is cleaved off. Measurement of the corresponding fluorescence of the dye indicates which base was incorporated. The process is repeated.

In another embodiment, the present method comprises obtaining sequence information for the nucleic acids in the test sample *e.g.* cfDNA in a maternal test sample, using nanopore sequencing (*e.g.* as described in Soni GV and Meller A. Clin Chem 53: 1996-2001 [2007]). Nanopore sequencing DNA analysis techniques are being industrially developed by a number of companies, including Oxford Nanopore Technologies (Oxford, United Kingdom). Nanopore sequencing is a single-molecule sequencing technology whereby a single molecule of DNA is sequenced directly as it passes through a nanopore. A nanopore is a small hole, of the order of 1 nanometer in diameter. Immersion of a nanopore in a conducting fluid and application of a potential (voltage) across it results in a slight electrical current due to conduction of ions through the nanopore. The amount of current which flows is sensitive to the size and shape of the nanopore. As a DNA molecule passes through a nanopore, each nucleotide on the DNA molecule obstructs the nanopore to a different degree, changing the magnitude of the current through the nanopore in different degrees. Thus, this change in the current as the DNA molecule passes through the nanopore represents a reading of the DNA sequence.

In another embodiment, the present method comprises obtaining sequence information for the nucleic acids in the test sample *e.g.* cfDNA in a maternal test sample, using the chemical-sensitive field effect transistor (chemFET) array (e.g., as described in U.S. Patent Application Publication No. 20090026082). In one example of the technique, DNA molecules can be placed into reaction chambers, and the template molecules can be hybridized to a sequencing primer bound to a polymerase. Incorporation of one or more triphosphates into a new nucleic acid strand at the 3' end of the sequencing primer can be discerned by a change in current by a chemFET. An array can have multiple chemFET sensors. In another example, single nucleic acids can be attached to beads, and the nucleic acids can be amplified on the bead, and the individual beads can be transferred to individual reaction chambers on a chemFET array, with each chamber having a chemFET sensor, and the nucleic acids can be sequenced.

In another embodiment, the present method comprises obtaining sequence information for the nucleic acids in the test sample *e.g.* cfDNA in a maternal test sample, using the Halcyon Molecular's technology, which uses transmission electron microscopy (TEM). The method, termed Individual Molecule Placement Rapid Nano Transfer (IMPRNT), comprises utilizing single atom resolution transmission electron microscope imaging of high-molecular weight (150kb or greater) DNA selectively labeled with heavy atom markers and arranging these molecules on ultra-thin films in ultra-dense (3nm strand-to-strand) parallel arrays with consistent base-to-base spacing. The electron microscope is used to image the molecules on the films to determine the position of the heavy atom markers and to extract base sequence information from the DNA. The method is further described in PCT patent publication WO 2009/046445. The method allows for sequencing complete human genomes in less than ten minutes.

In another embodiment, the DNA sequencing technology is the Ion Torrent single molecule sequencing, which pairs semiconductor technology with a simple sequencing chemistry to directly translate chemically encoded information (A, C, G, T) into digital information (0, 1) on a semiconductor chip. In nature, when a nucleotide is incorporated into a strand of DNA by a polymerase, a hydrogen ion is released as a byproduct. Ion Torrent uses a high-density array of micro-machined wells to perform this biochemical process in a massively parallel way. Each well holds a different DNA molecule. Beneath the wells is an ion-sensitive layer and beneath that an ion sensor. When a nucleotide, for example a C, is added to a DNA template and is then incorporated into a strand of DNA, a hydrogen ion will be released. The charge from that ion will change the pH of the solution, which can be detected by Ion Torrent's ion sensor. The sequencer—essentially the world's smallest solid-state pH meter—calls the base, going directly from chemical information to digital information. The Ion personal Genome Machine (PGM™) sequencer then sequentially floods the chip with one nucleotide after another. If the next nucleotide that floods the chip is not a match. No voltage change will be recorded and no base will be called. If there are two identical bases on the

DNA strand, the voltage will be double, and the chip will record two identical bases called. Direct detection allows recordation of nucleotide incorporation in seconds.

In another embodiment, the present method comprises obtaining sequence information for the nucleic acids in the test sample *e.g.* cfDNA in a maternal test sample, using sequencing by hybridization., Seqeuncing-by-hybridization comprises contacting the plurality of polynucleotide sequences with a plurality of polynucleotide probes, wherein each of the plurality of polynucleotide probes can be optionally tethered to a substrate. The substrate might be flat surface comprising an array of known nucleotide sequences. The pattern of hybridization to the array can be used to determine the polynucleotide sequences present in the sample. In other embodiments, each probe is tethered to a bead, e.g., a magnetic bead or the like. Hybridization to the beads can be determined and used to identify the plurality of polynucleotide sequences within the sample.

In another embodiment, the present method comprises obtaining sequence information for the nucleic acids in the test sample *e.g.* cfDNA in a maternal test sample, by massively parallel sequencing of millions of DNA fragments using Illumina's sequencing-by-synthesis and reversible terminator-based sequencing chemistry (*e.g.* as described in Bentley *et al.,* Nature 6:53-59 [2009]). Template DNA can be genomic DNA *e.g.* cfDNA. In some embodiments, genomic DNA from isolated cells is used as the template, and it is fragmented into lengths of several hundred base pairs. In other embodiments, cfDNA is used as the template, and fragmentation is not required as cfDNA exists as short fragments. For example fetal cfDNA circulates in the bloodstream as fragments approximately 170 base pairs (bp) in length (Fan *et al.,* Clin Chem 56:1279-1286 [2010]), and no fragmentation of the DNA is required prior to sequencing. Illumina's sequencing technology relies on the attachment of fragmented genomic DNA to a planar, optically transparent surface on which oligonucleotide anchors are bound. Template DNA is end-repaired to generate 5'-phosphorylated blunt ends, and the polymerase activity of Klenow fragment is used to add a single A base to the 3' end of the blunt phosphorylated DNA fragments. This addition prepares the DNA fragments for ligation to oligonucleotide adapters, which have an overhang of a single T base at their 3' end to increase ligation efficiency. The adapter oligonucleotides are complementary to the flow-cell anchors. Under limiting-dilution conditions, adapter-modified, single-stranded template DNA is added to the flow cell and immobilized by hybridization to the anchors. Attached DNA fragments are extended and bridge amplified to create an ultra-high density sequencing flow cell with hundreds of millions of clusters, each containing ~1,000 copies of the same template. In one embodiment, the randomly fragmented genomic DNA *e.g.* cfDNA, is amplified using PCR before it is subjected to cluster amplification. Alternatively, an amplification-free genomic library preparation is used, and the randomly fragmented genomic DNA *e.g.* cfDNA is enriched using the cluster amplification alone (Kozarewa *et al.,* Nature Methods 6:291-295 [2009]). The templates are sequenced using a robust four-color DNA sequencing-by-synthesis technology that employs reversible terminators with removable fluorescent

dyes. High-sensitivity fluorescence detection is achieved using laser excitation and total internal reflection optics. Short sequence reads of about 20-40 bp *e.g.* 36 bp, are aligned against a repeat-masked reference genome and unique mapping of the short sequence reads to the reference genome are identified using specially developed data analysis pipeline software. Non-repeat-masked reference genomes can also be used. Whether repeat-masked or non-repeat-masked reference genomes are used, only reads that map uniquely to the reference genome are counted. After completion of the first read, the templates can be regenerated in situ to enable a second read from the opposite end of the fragments. Thus, either single-end or paired end sequencing of the DNA fragments can be used. Partial sequencing of DNA fragments present in the sample is performed, and sequence tags comprising reads of predetermined length *e.g.* 36 bp, are mapped to a known reference genome are counted. In one embodiment, the reference genome sequence is the NCBI36/hg18 sequence, which is available on the world wide web at genome.ucsc.edu/cgi-bin/hgGateway?org=Human&db=hg18&hgsid=166260105). Alternatively, the reference genome sequence is the GRCh37/hg19, which is available on the world wide web at genome.ucsc.edu/cgi-bin/hgGateway. Other sources of public sequence information include GenBank, dbEST, dbSTS, EMBL (the European Molecular Biology Laboratory), and the DDBJ (the DNA Databank of Japan). A number of computer algorithms are available for aligning sequences, including without limitation BLAST (Altschul et al., 1990), BLITZ (MPsrch) (Sturrock & Collins, 1993), FASTA (Person & Lipman, 1988), BOWTIE (Langmead *et al.,* Genome Biology 10:R25.1-R25.10 [2009]), or ELAND (Illumina, Inc., San Diego, CA, USA). In one embodiment, one end of the clonally expanded copies of the plasma cfDNA molecules is sequenced and processed by bioinformatic alignment analysis for the Illumina Genome Analyzer, which uses the Efficient Large-Scale Alignment of Nucleotide Databases (ELAND) software.

In some embodiments of the method described herein, the mapped sequence tags comprise sequence reads of about 20bp, about 25bp, about 30bp, about 35bp, about 40bp, about 45bp, about 50bp, about 55bp, about 60bp, about 65bp, about 70bp, about 75bp, about 80bp, about 85bp, about90bp, about 95bp, about 100bp, about 110bp, about 120bp, about 130, about 140bp, about 150bp, about 200bp, about 250bp, about 300bp, about 350bp, about 400bp, about 450bp, or about 500bp. It is expected that technological advances will enable single-end reads of greater than 500bp enabling for reads of greater than about 1000bp when paired end reads are generated. In one embodiment, the mapped sequence tags comprise sequence reads that are 36bp. Mapping of the sequence tags is achieved by comparing the sequence of the tag with the sequence of the reference to determine the chromosomal origin of the sequenced nucleic acid (*e.g.* cfDNA) molecule, and specific genetic sequence information is not needed. A small degree of mismatch (0-2 mismatches per sequence tag) may be allowed to account for minor polymorphisms that may exist between the reference genome and the genomes in the mixed sample.

A plurality of sequence tags are obtained per sample. In some embodiments, at least about 3 x 10^6 sequence tags, at least about 5 x 10^6 sequence tags, at least about 8 x 10^6 sequence tags, at least about 10 x 10^6 sequence tags, at least about 15 x 10^6 sequence tags, at least about 20 x 10^6 sequence tags, at least about 30 x 10^6 sequence tags, at least about 40 x 10^6 sequence tags, or at least about 50 x 10^6 sequence tags comprising between 20 and 40bp reads *e.g.* 36bp, are obtained from mapping the reads to the reference genome per sample. In one embodiment, all the sequence reads are mapped to all regions of the reference genome. In one embodiment, the tags that have been mapped to all regions *e.g.* all chromosomes, of the reference genome are counted, and the CNV *i.e.* the over- or under-representation of a sequence of interest *e.g.* a chromosome or portion thereof, in the mixed DNA sample is determined. The method does not require differentiation between the two genomes.

The accuracy required for correctly determining whether a CNV *e.g.* aneuploidy, is present or absent in a sample, is predicated on the variation of the number of sequence tags that map to the reference genome among samples within a sequencing run (inter-chromosomal variability), and the variation of the number of sequence tags that map to the reference genome in different sequencing runs (inter-sequencing variability). For example, the variations can be particularly pronounced for tags that map to GC-rich or GC-poor reference sequences. Other variations can result from using different protocols for the extraction and purification of the nucleic acids, the preparation of the sequencing libraries, and the use of different sequencing platforms. The present method uses sequence doses (chromosome doses, or segment doses) based on the knowledge of normalizing sequences (normalizing chromosome sequences or normalizing segment sequences), to intrinsically account for the accrued variability stemming from interchromosomal (intra-run), and inter-sequencing (inter-run) and platform-dependent variability. Chromosome doses are based on the knowledge of a normalizing chromosome sequence, which can be composed of a single chromosome, or of two or more chromosomes selected from chromosomes 1-22, X, and Y. Alternatively, normalizing chromosome sequences can be composed of a single chromosome segment, or of two or more segments of one chromosome or of two or more chromosomes. Segment doses are based on the knowledge of a normalizing segment sequence, which can be composed of a single segment of any one chromosome, or of two or more segments of any two or more of chromosomes 1-22, X, and Y.

Determination of Normalizing Sequences in Qualified Samples: Normalizing Chromosome Sequences and Normalizing Segment Sequences

Normalizing sequences are identified using sequence information from a set of qualified samples obtained from subjects known to comprise cells having a normal copy number for any one sequence of interest *e.g.* a chromosome or segment thereof. Determination of normalizing sequences is outlined in steps **100**, **120**, **130**, **140**, and **145** of the embodiment of the method depicted in **Figure 1**. The sequence information obtained from the qualified samples is also used for determining

statistically meaningful identification of chromosomal aneuploidies in test samples (step **155 Figure 1**, and Examples).

Figure 1 provides a flow diagram of an embodiment of the method of the invention **100** for determining a CNV of a sequence of interest *e.g.* a chromosome or segment thereof, in a biological sample. In some embodiments, a biological sample is obtained from a subject and comprises a mixture of nucleic acids contributed by different genomes. The different genomes can be contributed to the sample by two individuals *e.g.* the different genomes are contributed by the fetus and the mother carrying the fetus. Alternatively, the genomes are contributed to the sample by aneuploid cancerous cells and normal euploid cells from the same subject *e.g.* a plasma sample from a cancer patient.

A set of qualified samples is obtained to identify qualified normalizing sequences and to provide variance values for use in determining statistically meaningful identification of CNV in test samples. In step **110**, a plurality of biological qualified samples are obtained from a plurality of subjects known to comprise cells having a normal copy number for any one sequence of interest. In one embodiment, the qualified samples are obtained from mothers pregnant with a fetus that has been confirmed using cytogenetic means to have a normal copy number of chromosomes. The biological qualified samples may be a biological fluid *e.g.* plasma, or any suitable sample as described below. In some embodiments, a qualified sample contains a mixture of nucleic acid molecules *e.g.* cfDNA molecules. In some embodiments, the qualified sample is a maternal plasma sample that contains a mixture of fetal and maternal cfDNA molecules. Sequence information for normalizing chromosomes and/or segments thereof is obtained by sequencing at least a portion of the nucleic acids *e.g.* fetal and maternal nucleic acids, using any known sequencing method. Preferably, any one of the Next Generation Sequencing (NGS) methods described elsewhere herein is used to sequence the fetal and maternal nucleic acids as single or clonally amplified molecules.

In step **120**, at least a portion of each of all the qualified nucleic acids contained in the qualified samples are sequenced to generate millions of sequence reads *e.g.* 36bp reads, which are aligned to a reference genome, *e.g.* hg18. In some embodiments, the sequence reads comprise about 20bp, about 25bp, about 30bp, about 35bp, about 40bp, about 45bp, about 50bp, about 55bp, about 60bp, about 65bp, about 70bp, about 75bp, about 80bp, about 85bp, about90bp, about 95bp, about 100bp, about 110bp, about 120bp, about 130, about 140bp, about 150bp, about 200bp, about 250bp, about 300bp, about 350bp, about 400bp, about 450bp, or about 500bp. It is expected that technological advances will enable single-end reads of greater than 500bp enabling for reads of greater than about 1000bp when paired end reads are generated. In one embodiment, the mapped sequence reads comprise 36bp. Sequence reads are aligned to a reference genome, and the reads that are uniquely mapped to the reference genome are known as sequence tags. In one embodiment, at least about 3×10^6 qualified sequence tags, at least about 5×10^6 qualified sequence tags, at least about

8×10^6 qualified sequence tags, at least about 10×10^6 qualified sequence tags, at least about 15×10^6 qualified sequence tags, at least about 20×10^6 qualified sequence tags, at least about 30×10^6 qualified sequence tags, at least about 40×10^6 qualified sequence tags, or at least about 50×10^6 qualified sequence tags comprising between 20 and 40bp reads are obtained from reads that map uniquely to a reference genome.

In step **130**, all the tags obtained from sequencing the nucleic acids in the qualified samples are counted to determine a qualified sequence tag density. In one embodiment the sequence tag density is determined as the number of qualified sequence tags mapped to the sequence of interest on the reference genome. In another embodiment, the qualified sequence tag density is determined as the number of qualified sequence tags mapped to a sequence of interest normalized to the length of the qualified sequence of interest to which they are mapped. Sequence tag densities that are determined as a ratio of the tag density relative to the length of the sequence of interest are herein referred to as tag density ratios. Normalization to the length of the sequence of interest is not required, and may be included as a step to reduce the number of digits in a number to simplify it for human interpretation. As all qualified sequence tags are mapped and counted in each of the qualified samples, the sequence tag density for a sequence of interest *e.g.* a clinically-relevant sequence, in the qualified samples is determined, as are the sequence tag densities for additional sequences from which normalizing sequences are identified subsequently.

In some embodiments, the sequence of interest is a chromosome that is associated with a complete chromosomal aneuploidy *e.g.* chromosome 21, and the qualified normalizing sequence is a complete chromosome that is not associated with a chromosomal aneuploidy and whose variation in sequence tag density best approximates that of the sequence (*i.e.* chromosome) of interest *e.g.* chromosome 21. Any one or more of chromosomes 1-22, X, and Y can be a sequence of interest, and one or more chromosomes can be identified as the normalizing sequence for each of the any one chromosomes 1-22, X and Y in the qualified samples. The normalizing chromosome can be an individual chromosome or it can be a group of chromosomes as described elsewhere herein.

In another embodiment, the sequence of interest is a segment of a chromosome associated with a partial aneuploidy, *e.g.* a chromosomal deletion or insertion, or unbalanced chromosomal translocation, and the normalizing sequence is a chromosomal segment that is not associated with the partial aneuploidy and whose variation in sequence tag density best approximates that of the chromosome segment associated with the partial aneuploidy. Any one or more segments of any one or more chromosomes 1-22, X, and Y can be a sequence of interest.

In all embodiments, whether a single sequence or a group of sequences are identified in the qualified samples as the normalizing sequence for any one or more sequence of interest, the qualified normalizing sequence has a variation in sequence tag density best approximates that of the sequence of interest as determined in the qualified samples. For example, a qualified normalizing sequence is a

sequence that has the smallest variability *i.e.* the variability of the normalizing sequence is closest to that of the sequence of interest.

In some embodiments, the normalizing sequence is a sequence that best distinguishes one or more qualified, samples from one or more affected samples, which implies that the normalizing sequence is a sequence that has the greatest differentiability *i.e.* the differentiability of the normalizing sequence is such that it provides optimal differentiation to a sequence of interest in an affected test sample to easily distinguish the affected test sample from other unaffected samples. In other embodiments, the normalizing sequence is a sequence that has the smallest variability and the greatest differentiability. The level of differentiability can be determined as a statistical difference between the sequence doses *e.g.* chromosome doses or segment doses, in a population of qualified samples and the chromosome dose(s) in one or more test samples as described below and shown in the Examples. For example, differentiability can be represented numerically as a T-test value, which represents the statistical difference between the chromosome doses in a population of qualified samples and the chromosome dose(s) in one or more test samples. Alternatively, differentiability can be represented numerically as a Normalized Chromosome Value (NCV), which is a z-score for chromosome doses as long as the distribution for the NCV is normal. Similarly, differentiability can be represented numerically as a T-test value, which represents the statistical difference between the segment doses in a population of qualified samples and the segment dose(s) in one or more test samples. Alternatively, differentiability of segment doses can be represented numerically as a Normalized Segment Value (NSV), which is a z-score for chromosome doses as long as the distribution for the NSV is normal. In determining the z-score, the mean and standard deviation of chromosome or segment doses in a set of qualified samples can be used. Alternatively, the mean and standard deviation of chromosome or segment doses in a training set comprising qualified samples and affected samples can be used. In other embodiments, the normalizing sequence is a sequence that has the smallest variability and the greatest differentiability.

The method identifies sequences that inherently have similar characteristics and that are prone to similar variations among samples and sequencing runs, and which are useful for determining sequence doses in test samples.

Determination of sequence doses (*i.e.* chromosome doses or segment doses) in qualified samples

In step **140**, based on the calculated qualified tag densities, a qualified sequence dose *i.e.* a chromosome dose or a segment dose, for a sequence of interest is determined as the ratio of the sequence tag density for the sequence of interest and the qualified sequence tag density for additional sequences from which normalizing sequences are identified subsequently in step 145. The identified normalizing sequences are used subsequently to determine sequence doses in test samples.

In one embodiment, the sequence dose in the qualified samples is a chromosome dose that is calculated as the ratio of the number of sequence tags for a chromosome of interest and the number of

sequence tags for a normalizing chromosome sequence in a qualified sample. The normalizing chromosome sequence can be a single chromosome, a group of chromosomes, a segment of one chromosome, or a group of segments from different chromosomes. Accordingly, a chromosome dose for a chromosome of interest is determined in a qualified sample as (i) the ratio of the number of tags for a chromosome of interest and the number of tags for a normalizing chromosome sequence composed of a single chromosome, (ii) the ratio of the number of tags for a chromosome of interest and the number of tags for a normalizing chromosome sequence composed of two or more chromosomes, or (iii) the ratio of the number of tags for a chromosome of interest and the number of tags for a normalizing segment sequence composed of a single segment of a chromosome, (iv) the ratio of the number of tags for a chromosome of interest and the number of tags for a normalizing segment sequence composed of two or more segments form one chromosome, or (v) the ratio of the number of tags for a chromosome of interest and the number of tags for a normalizing segment sequence composed of two or more segments of two or more chromosomes. Examples for determining a chromosome dose for chromosome of interest 21 according to (i)-(v) are as follows: chromosome doses for chromosome of interest *e.g.* chromosome 21, are determined as a ratio of the sequence tag density of chromosome 21 and the sequence tag density for each of all the remaining chromosomes *i.e.* chromosomes 1-20, chromosome 22, chromosome X, and chromosome Y (i); chromosome doses for chromosome of interest *e.g.* chromosome 21, are determined as a ratio of the sequence tag density of chromosome 21 and the sequence tag density for all possible combinations of two or more remaining chromosomes (ii); chromosome doses for chromosome of interest *e.g.* chromosome 21, are determined as a ratio of the sequence tag density of chromosome 21 and the sequence tag density for a segment of another chromosome *e.g.* chromosome 9 (iii); chromosome doses for chromosome of interest *e.g.* chromosome 21, are determined as a ratio of the sequence tag density of chromosome 21 and the sequence tag density for two segment of one another chromosome *e.g.* two segments of chromosome 9 (iv); and chromosome doses for chromosome of interest *e.g.* chromosome 21, are determined as a ratio of the sequence tag density of chromosome 21 and the sequence tag density for two segments of two different chromosomes *e.g.* a segment of chromosome 9 and a segment of chromosome 14.

In another embodiment, the sequence dose in the qualified samples is a segment dose that is calculated as the ratio of the number of sequence tags for a segment of interest and the number of sequence tags for a normalizing segment sequence in a qualified sample. The normalizing segment sequence can be a segment of one chromosome, or a group of segments from different chromosomes. Accordingly, a segment dose for a segment of interest is determined in a qualified sample as (i) the ratio of the number of tags for a segment of interest and the number of tags for a normalizing segment sequence composed of a single segment of a chromosome, (ii) the ratio of the number of tags for a segment of interest and the number of tags for a normalizing segment sequence composed of two or

more segments of one chromosome, or (iii) the ratio of the number of tags for a segment of interest and the number of tags for a normalizing segment sequence composed of two or more segments of two or more different chromosomes.

Chromosome doses for one or more chromosomes of interest are determined in all qualified samples, and a normalizing chromosome sequence is identified in step **145**. Similarly, segment doses for one or more segments of interest are determined in all qualified samples, and a normalizing segment sequence is identified in step **145.**

Identification of normalizing sequences from qualified sequence doses

In step **145**, a normalizing sequence is identified for a sequence of interest as the sequence based on the calculated sequence doses *i.e.* that results in the smallest variability in sequence dose for the sequence of interest across all qualified samples. The method identifies sequences that inherently have similar characteristics and that are prone to similar variations among samples and sequencing runs, and which are useful for determining sequence doses in test samples.

Normalizing sequences for one or more sequences of interest can be identified in a set of qualified samples, and the sequences that are identified in the qualified samples are used subsequently to calculate sequence doses for one or more sequences of interest in each of the test samples (step **150**) to determine the presence or absence of aneuploidy in each of the test samples. The normalizing sequence identified for chromosomes or segments of interest may differ when different sequencing platforms are used and/or when differences exist in the purification of the nucleic acid that is to be sequenced and/or preparation of the sequencing library. The use of normalizing sequences according to the method of the invention provides specific and sensitive measure of a variation in copy number of a chromosome or segment thereof irrespective of sample preparation and/or sequencing platform that is used.

In some embodiments, more than one normalizing sequence is identified *i.e.* different normalizing sequences can be determined for one sequence of interest, and multiple sequence doses can be determined for one sequence of interest. For example, the variation *e.g.* coefficient of variation, in chromosome dose for chromosome of interest 21 is least when the sequence tag density of chromosome 14 is used. However, two, three, four, five, six, seven, eight or more normalizing sequences can be identified for use in determining a sequence dose for a sequence of interest in a test sample. As an example, a second dose for chromosome 21 in any one test sample can be determined using chromosome 7, chromosome 9, chromosome 11 or chromosome 12 as the normalizing chromosome sequence as these chromosomes all have CV close to that for chromosome 14 (see Example 2, Table 2). Preferably, when a single chromosome is chosen as the normalizing chromosome sequence for a chromosome of interest, the normalizing chromosome sequence will be a chromosome that results in chromosome doses for the chromosome of interest that has the smallest variability across all samples tested *e.g.* qualified samples.

Normalizing chromosome sequence as a normalizing sequence for chromosome(s)

In other embodiments, a normalizing chromosome sequence can be a single sequence or it can be a group of sequences. For example, in some embodiments, a normalizing sequence is a group of sequences *e.g.* a group of chromosomes, that is identified as the normalizing sequence for any or more of chromosomes 1-22, X and Y. The group of chromosomes that compose the normalizing sequence for a chromosome of interest *i.e.* a normalizing chromosome sequence, can be a group of two, three, four, five, six, seven, eight, nine, ten, eleven, twelve, thirteen, fourteen, fifteen, sixteen, seventeen, eighteen, nineteen, twenty, twenty-one, or twenty-two chromosomes, and including or excluding one or both of chromosomes X , and Y. The group of chromosomes that is identified as the normalizing chromosome sequence is a group of chromosomes that results in chromosome doses for the chromosome of interest that has the smallest variability across all samples tested *e.g.* qualified samples. Preferably, individual and groups of chromosomes are tested together for their ability to best mimic the behavior of the sequence of interest for which they are chosen as normalizing chromosome sequences.

In one embodiment, the normalizing sequence for chromosome 21 is selected from chromosome 9, chromosome 1, chromosome 2, chromosome 3, chromosome 4, chromosome 5, chromosome 6, chromosome 7, chromosome 8, chromosome 10, chromosome 11, chromosome 12, chromosome 13, chromosome 14, chromosome 15, chromosome 16, and chromosome 17. In another embodiment, the normalizing sequence for chromosome 21 is selected from chromosome 9, chromosome 1, chromosome 2, chromosome 11, chromosome 12, and chromosome 14. Alternatively, the normalizing sequence for chromosome 21 is a group of chromosomes selected from chromosome 9, chromosome 1, chromosome 2, chromosome 3, chromosome 4, chromosome 5, chromosome 6, chromosome 7, chromosome 8, chromosome 10, chromosome 11, chromosome 12, chromosome 13, chromosome 14, chromosome 15, chromosome 16, and chromosome 17. In another embodiment, the group of chromosomes is a group selected from chromosome 9, chromosome 1, chromosome 2, chromosome 11, chromosome 12, and chromosome 14.

In some embodiments the method is further improved by using a normalizing sequence that is determined by systematic calculation of all chromosome doses using each chromosome individually and in all possible combinations with all remaining chromosomes (see Example 7). For example, a systematically determined normalizing chromosome can be determined for each chromosome of interest by systematically calculating all possible chromosome doses using one of any of chromosomes 1-22, X, and Y, and combinations of two or more of chromosomes 1-22, X, and Y to determine which single or group of chromosomes is the normalizing chromosome that results in the least variability of the chromosome dose for a chromosome of interest across a set of qualified samples (see Example 7). Accordingly, in one embodiment, the systematically calculated normalizing chromosome sequence for chromosome 21 is a group of chromosomes consisting of chromosome 4,

chromosome 14, chromosome 16, chromosome 20,-and chromosome 22. Single or groups of chromosomes can be determined for all chromosomes in the genome.

In one embodiment, the normalizing sequence for chromosome 18 is selected chromosome 8, chromosome 2, chromosome 3, chromosome 4, chromosome 5, chromosome 6, chromosome 7, chromosome 9, chromosome 10, chromosome 11, chromosome 12, chromosome 13, and chromosome 14. Preferably, the normalizing sequence for chromosome 18 is selected from chromosome 8, chromosome 2, chromosome 3, chromosome 5, chromosome 6, chromosome 12, and chromosome 14. Alternatively, the normalizing sequence for chromosome 18 is a group of chromosomes selected from chromosome 8, chromosome 2, chromosome 3, chromosome 4, chromosome 5, chromosome 6, chromosome 7, chromosome 9, chromosome 10, chromosome 11, chromosome 12, chromosome 13, and chromosome 14. Preferably, the group of chromosomes is a group selected from chromosome 8, chromosome 2, chromosome 3, chromosome 5, chromosome 6, chromosome 12, and chromosome 14.

In another embodiment, the normalizing sequence for chromosome 18 is determined by systematic calculation of all possible chromosome doses using each possible normalizing chromosome individually and all possible combinations of normalizing chromosomes (as explained elsewhere herein). Accordingly, in one embodiment, the normalizing sequence for chromosome 18 is a normalizing chromosome consisting of the group of chromosomes consisting of chromosome 2, chromosome 3, chromosome 5, and chromosome 7.

In one embodiment, the normalizing sequence for chromosome X is selected from chromosome 1, chromosome 2, chromosome 3, chromosome 4, chromosome 5, chromosome 6, chromosome 7, chromosome 8, chromosome 9, chromosome 10, chromosome 11, chromosome 12, chromosome 13, chromosome 14, chromosome 15, and chromosome 16. Preferably, the normalizing sequence for chromosome X is selected from chromosome 2, chromosome 3, chromosome 4, chromosome 5, chromosome 6 and chromosome 8. Alternatively, the normalizing sequence for chromosome X is a group of chromosomes selected from chromosome 1, chromosome 2, chromosome 3, chromosome 4, chromosome 5, chromosome 6, chromosome 7, chromosome 8, chromosome 9, chromosome 10, chromosome 11, chromosome 12, chromosome 13, chromosome 14, chromosome 15, and chromosome 16. Preferably, the group of chromosomes is a group selected from chromosome 2, chromosome 3, chromosome 4, chromosome 5, chromosome 6, and chromosome 8.

In another embodiment, the normalizing sequence for chromosome X is determined by systematic calculation of all possible chromosome doses using each possible normalizing chromosome individually and all possible combinations of normalizing chromosomes (as explained elsewhere herein). Accordingly, in one embodiment, the normalizing sequence for chromosome X is a normalizing chromosome consisting of the group of chromosome 4 and chromosome 8.

In one embodiment, the normalizing sequence for chromosome 13 is a chromosome selected from chromosome 2, chromosome 3, chromosome 4, chromosome 5, chromosome 6, chromosome 7,

chromosome 8, chromosome 9, chromosome 10, chromosome 11, chromosome 12, chromosome 14, chromosome 18, and chromosome 21. Preferably, the normalizing sequence for chromosome 13 is a chromosome selected from chromosome 2, chromosome 3, chromosome 4, chromosome 5, chromosome 6, and chromosome 8. In another embodiment, the normalizing sequence for chromosome 13 is a group of chromosomes selected from chromosome 2, chromosome 3, chromosome 4, chromosome 5, chromosome 6, chromosome 7, chromosome 8, chromosome 9, chromosome 10, chromosome 11, chromosome 12, chromosome 14, chromosome 18, and chromosome 21. Preferably, the group of chromosomes is a group selected from chromosome 2, chromosome 3, chromosome 4, chromosome 5, chromosome 6, and chromosome 8.

In another embodiment, the normalizing sequence for chromosome 13 is determined by systematic calculation of all possible chromosome doses using each possible normalizing chromosome individually and all possible combinations of normalizing chromosomes (as explained elsewhere herein). Accordingly, in one embodiment, the normalizing sequence for chromosome 13 is a normalizing chromosome comprising the group of chromosome 4 and chromosome 5. In another embodiment, the normalizing sequence for chromosome 13 is a normalizing chromosome consisting of the group of chromosome 4 and chromosome 5.

The variation in chromosome dose for chromosome Y is greater than 30 independently of which normalizing chromosome is used in determining the chromosome Y dose. Therefore, any one chromosome, or a group of two or more chromosomes selected from chromosomes 1-22 and chromosome X can be used as the normalizing sequence for chromosome Y. In one embodiment, the at least one normalizing chromosome is a group of chromosomes consisting of chromosomes 1-22, and chromosome X. In another embodiment, the group of chromosomes consists of chromosome 2, chromosome 3, chromosome 4, chromosome 5, and chromosome 6.

In another embodiment, the normalizing sequence for chromosome Y is determined by systematic calculation of all possible chromosome doses using each possible normalizing chromosome individually and all possible combinations of normalizing chromosomes (as explained elsewhere herein). Accordingly, in one embodiment, the normalizing sequence for chromosome Y is a normalizing chromosome comprising the group of chromosomes consisting of chromosome 4 and chromosome 6. In another embodiment, the normalizing sequence for chromosome Y is a normalizing chromosome consisting of the group of chromosomes consisting of chromosome 4 and chromosome 6.

The normalizing sequence used to calculate the dose of different chromosomes of interest, or of different segments of interest can be the same or it can be a different normalizing sequence for different chromosomes or segments of interest, respectively. For example, the normalizing sequence *e.g.* a normalizing chromosome (one or a group) for chromosome of interest A can be the same or it

can be different from the normalizing sequence *e.g.* a normalizing chromosome (one or a group) for chromosome of interest B.

The normalizing sequence for a complete chromosome may be a complete chromosome or a group of complete chromosomes, or it may be a segment of a chromosome, or a group of segments of one or more chromosomes.

Normalizing segment sequence as a normalizing sequence for chromosome(s)

In another embodiment, the normalizing sequence for a chromosome can be a normalizing segment sequence. The normalizing segment sequence can be a single segment or it can be a group of segments of one chromosome, or they can be segments from two or more different chromosomes. A normalizing segment sequence can be determined by systematic calculation of all combinations of segment sequences in the genome. For example, a normalizing segment sequence for chromosome 21 can be a single segment that is bigger or smaller than the size of chromosome 2, which is approximately 47Mbp (million base pairs) from chromosome 9, which is approximately 140 Mbp. Alternatively, a normalizing sequence for chromosome 21 can be a combination of a sequence form chromosome 1, and a sequence from chromosome 12.

In one embodiment, the normalizing sequence for chromosome 21 is a normalizing segment sequence of one segment or of a group of two or more segments of chromosomes 1-20, 22, X, and Y. In another embodiment, the normalizing sequence for chromosome 18 is a segment or groups segments of chromosomes 1–17, 19-22, X, and Y. In another embodiment, the normalizing sequence for chromosome 13 is a segment or groups of segments of chromosomes 1-,12, 14-22, X, and Y. In another embodiment, the normalizing sequence for chromosome X is a segment or groups segments of chromosomes 1-22, and Y. In another embodiment, the normalizing sequence for chromosome Y is a segment or group of segments of chromosomes 1-,22, and X. Normalizing segment sequences of single or groups of segments can be determined for all chromosomes in the genome. The two or more segments of a normalizing segment sequence can be segments from one chromosome, or the two or more segments can be segments of two or more different chromosomes. As described for normalizing chromosome sequences, a normalizing segment sequence can be the same for two or more different chromosomes.

Normalizing segment sequence as a normalizing sequence for chromosome segment(s)

The presence or absence of CNV of a sequence of interest can be determined when the sequence of interest is a segment of a chromosome. Variation in the copy number of a chromosome segment allows for determining the presence or absence of a partial chromosomal aneuploidy. Described below are examples of partial chromosomal aneuploidies that are associated with various fetal abnormalities and disease conditions. The segment of the chromosome can be of any length. For example, it can range from a kilobase to hundreds of megabases. The human genome occupies just over 3 billion DNA bases, which can be divided into tens, thousands, hundreds of thousands and

millions of segments of different sizes of which the copy number can be determined according to the present method. The normalizing sequence for a segment of a chromosome is a normalizing segment sequence, which can be a single segment from any one of the chromosomes 1-22, X and Y, or it can be a group of segments from any one or more of chromosomes 1-22, X, and Y.

The normalizing sequence for a segment of interest is a sequence that has a variability across chromosomes and across samples that is closest to that of the segment of interest. Determination of a normalizing sequence can be performed as described for determining the normalizing sequence for a chromosome of interest when the normalizing sequence is a group of segments of any one or more of chromosomes 1-22, X and Y. A normalizing segment sequence of one or a group of segments can be identified by calculating segment doses using one, and all possible combinations of two or more segments as normalizing sequences for the segment of interest in each sample of a set of qualified samples *i.e.* samples known to be diploid for the segment of interest, and the normalizing sequence is determined as that providing a segment dose having the lowest variability for the segment of interest across all qualified samples, as is described above for normalizing chromosome sequences.

For example, for a segment of interest that is 1Mb (megabase), the remaining 3 million segments (minus the 1 mg segment of interest) of the approximately 3Gb human genome can be used individually or in combination with each other to calculate segment doses for a segment of interest in a qualified set of sample to determine which one or group of segments would serve as the normalizing segment sequence for qualified and test samples. Segments of interest can vary from about 1000 bases to tens of megabases. Normalizing segment sequences can be composed of one or more segments of the same size as that of the sequence of interest. In other embodiment, the normalizing segment sequence can be composed of segments that differ from that of the sequence of interest, and/or from each other. For example, a normalizing segment sequence for a 10,0000 base long sequence can be 20,000 bases long, and comprise a combination of sequences of different lengths *e.g.* a 7,000+8,000+5,000 bases. As is described elsewhere herein for normalizing chromosome sequences, normalizing segment sequences can be determined by systematic calculation of all possible chromosome and/or segment doses using each possible normalizing chromosome segment individually and all possible combinations of normalizing segments (as explained elsewhere herein). Single or groups of segments can be determined for all segments and/or chromosomes in the genome.

The normalizing sequence used to calculate the dose of different chromosome segments of interest can be the same or it can be a different normalizing sequence for different chromosome segments of interest. For example, the normalizing sequence *e.g.* a normalizing segment (one or a group) for chromosome segment of interest A can be the same or it can be different from the normalizing sequence *e.g.* a normalizing segment (one or a group) for chromosome segment of interest B.

Determination of aneuploidies in test samples

Based on the identification of the normalizing sequence(s) in qualified samples, a sequence dose is determined for a sequence of interest in a test sample comprising a mixture of nucleic acids derived from genomes that differ in one or more sequences of interest.

In step **115**, a test sample is obtained from a subject suspected or known to carry a clinically-relevant CNV of a sequence of interest. The test sample may be a biological fluid *e.g.* plasma, or any suitable sample as described below. In some embodiments, a test sample contains a mixture of nucleic acid molecules *e.g.* cfDNA molecules. In some embodiments, the test sample is a maternal plasma sample that contains a mixture of fetal and maternal cfDNA molecules.

In step **125**, at least a portion of the test nucleic acids in the test sample is sequenced as described for the qualified samples to generate millions of sequence reads *e.g.* 36bp reads. As in step **120**, the reads generated from sequencing the nucleic acids in the test sample are uniquely mapped to a reference genome. As described in step **120**, at least about 3 x 10^6 qualified sequence tags, at least about 5 x 10^6 qualified sequence tags, at least about 8 x 10^6 qualified sequence tags, at least about 10 x 10^6 qualified sequence tags, at least about 15 x 10^6 qualified sequence tags, at least about 20 x 10^6 qualified sequence tags, at least about 30 x 10^6 qualified sequence tags, at least about 40 x 10^6 qualified sequence tags, or at least about 50 x 10^6 qualified sequence tags comprising between 20 and 40bp reads are obtained from reads that map uniquely to a reference genome.

In step **135**, all the tags obtained from sequencing the nucleic acids in the test samples are counted to determine a test sequence tag density. In one embodiment, the number of test sequence tags mapped to a sequence of interest is normalized to the known length of a sequence of interest to which they are mapped to provide a test sequence tag density ratio. As described for the qualified samples, normalization to the known length of a sequence of interest is not required, and may be included as a step to reduce the number of digits in a number to simplify it for human interpretation. As all the mapped test sequence tags are counted in the test sample, the sequence tag density for a sequence of interest *e.g.* a clinically-relevant sequence, in the test samples is determined, as are the sequence tag densities for additional sequences that correspond to at least one normalizing sequence identified in the qualified samples.

In step **150**, based on the identity of at least one normalizing sequence in the qualified samples, a test sequence dose is determined for a sequence of interest in the test sample. As described elsewhere herein, the at least one normalizing sequence can be a single sequence or a group of sequences. The sequence dose for a sequence of interest in a test sample is a ratio of the sequence tag density determined for the sequence of interest in the test sample and the sequence tag density of at least one normalizing sequence determined in the test sample, wherein the normalizing sequence in the test sample corresponds to the normalizing sequence identified in the qualified samples for the particular sequence of interest. For example, if the normalizing sequence identified for chromosome

21 in the qualified samples is determined to be a chromosome *e.g.* chromosome 14, then the test sequence dose for chromosome 21 (sequence of interest) is determined as the ratio of the sequence tag density for chromosome 21 in and the sequence tag density for chromosome 14 each determined in the test sample. Similarly, chromosome doses for chromosomes 13, 18, X, Y, and other chromosomes associated with chromosomal aneuploidies are determined. A normalizing sequence for a chromosome of interest can be one or a group of chromosomes, or one or a group of chromosome segments. As described previously, a sequence of interest can be part of a chromosome *e.g.* a chromosome segment. Accordingly, the dose for a chromosome segment can be determined as the ratio of the sequence tag density determined for the segment in the test sample and the sequence tag density for the normalizing chromosome segment in the test sample, wherein the normalizing segment in the test sample corresponds to the normalizing segment (single or a group of segments) identified in the qualified samples for the particular segment of interest. Chromosome segments can range from kilobases (kb) to megabases (Mb) in size.

In step **155**, threshold values are derived from standard deviation values established for qualified sequence doses determined in a plurality of qualified samples and sequence doses determined for samples known to be aneuploid for a sequence of interest. Accurate classification depends on the differences between probability distributions for the different classes *i.e.* type of aneuploidy. Preferably, thresholds are chosen from empirical distribution for each type of aneuploidy *e.g.* trisomy 21. Possible threshold values that were established for classifying trisomy 13, trisomy 18, trisomy 21, and monosomy X aneuploidies as described in the Examples, which describe the use of the method for determining chromosomal aneuploidies by sequencing cfDNA extracted from a maternal sample comprising a mixture of fetal and maternal nucleic acids. The threshold value that is determined to distinguish samples affected for an aneuploidy of a chromosome can be the same or can be different from the threshold that is determined to distinguish samples affected for a different aneuploidy. As is shown in the Examples, the threshold value for each chromosome of interest is determined from the variability in the dose of the chromosome of interest across samples and sequencing runs. The less variable the chromosome dose for any chromosome of interest, the narrower the spread in the dose for the chromosome of interest across all the unaffected samples, which are used to set the threshold for determining different aneuploidies.

In step **160**, the copy number variation of the sequence of interest is determined in the test sample by comparing the test sequence dose for the sequence of interest to at least one threshold value established from the qualified sequence doses.

In step **165**, the calculated dose for a test sequence of interest is compared to that set as the threshold values that are chosen according to a user-defined threshold of reliability to classify the sample as a “normal” an “affected” or a “no call”. The “no call” samples are samples for which a definitive diagnosis cannot be made with reliability.

Another embodiment of the invention provides a method for providing prenatal diagnosis of a fetal chromosomal aneuploidy in a biological sample comprising fetal and maternal nucleic acid molecules. The diagnosis is made based on obtaining sequence information sequencing at least a portion of the mixture of the fetal and maternal nucleic acid molecules derived from a biological test sample *e.g.* a maternal plasma sample, computing from the sequencing data a normalizing chromosome dose for one or more chromosomes of interest, and/or a normalizing segment dose for one or more segments of interest, and determining a statistically significant difference between the chromosome dose for the chromosome of interest and/or the segment dose for the segment of interest, respectively, in the test sample and a threshold value established in a plurality of qualified (normal) samples, and providing the prenatal diagnosis based on the statistical difference. As described in step **165** of the method, a diagnosis of normal or affected is made. A “no call” is provided in the event that the diagnosis for normal or affected cannot be made with confidence.

Samples

Samples that are used for determining a CNV *e.g.* chromosomal and partial aneuploidies, comprise nucleic acids that are present in cells or that are “cell-free”. In some embodiments of the invention it is advantageous to obtain cell-free nucleic acids *e.g.* cell-free DNA (cfDNA). Cell-free nucleic acids, including cell-free DNA, can be obtained by various methods known in the art from biological samples including but not limited to plasma and serum (Chen *et al.,* Nature Med. 2: 1033-1035 [1996]; Lo *et al.,* Lancet 350: 485-487 [1997]). To separate cell-free DNA from cells, fractionation, centrifugation (e.g., density gradient centrifugation), DNA-specific precipitation, or high-throughput cell sorting and/or separation methods can be used.

The sample comprising the mixture of nucleic acids to which the methods described herein are applied is a biological sample such as a tissue sample, a biological fluid sample, or a cell sample. In some embodiments, the mixture of nucleic acids is purified or isolated from the biological sample by any one of the known methods. A sample can consist of purified or isolated polynucleotide, or it can comprise a biological sample such as a tissue sample, a biological fluid sample, or a cell sample. A biological fluid includes, as non-limiting examples, blood, plasma, serum, sweat, tears, sputum, urine, sputum, ear flow, lymph, saliva, cerebrospinal fluid, ravages, bone marrow suspension, vaginal flow, transcervical lavage, brain fluid, ascites, milk, secretions of the respiratory, intestinal and genitourinary tracts, amniotic fluid and leukophoresis samples. In some embodiments, the sample is a sample that is easily obtainable by non-invasive procedures *e.g.* blood, plasma, serum, sweat, tears, sputum, urine, sputum, ear flow, saliva or feces. Preferably, the biological sample is a peripheral blood sample, or the plasma and serum fractions. In other embodiments, the biological sample is a swab or smear, a biopsy specimen, or a cell culture. In another embodiment, the sample is a mixture of two or more biological samples *e.g.* a biological sample can comprise two or more of a biological fluid sample, a tissue sample, and a cell culture sample. As used herein, the terms "blood," "plasma"

and "serum" expressly encompass fractions or processed portions thereof. Similarly, where a sample is taken from a biopsy, swab, smear, *etc.*, the "sample" expressly encompasses a processed fraction or portion derived from the biopsy, swab, smear, *etc.*

In some embodiments, samples can be obtained from sources, including, but not limited to, samples from different individuals, different developmental stages of the same or different individuals, different diseased individuals (e.g., individuals with cancer or suspected of having a genetic disorder), normal individuals, samples obtained at different stages of a disease in an individual, samples obtained from an individual subjected to different treatments for a disease, samples from individuals subjected to different environmental factors, or individuals with predisposition to a pathology, or individuals with exposure to an infectious disease agent (e.g., HIV).

In one embodiment, the sample is a maternal sample that is obtained from a pregnant female, for example a pregnant woman. In this instance, the sample can be analyzed using the methods described herein to provide a prenatal diagnosis of potential chromosomal abnormalities in the fetus. The maternal sample can be a tissue sample, a biological fluid sample, or a cell sample. A biological fluid includes, as non-limiting examples, blood, plasma, serum, sweat, tears, sputum, urine, sputum, ear flow, lymph, saliva, cerebrospinal fluid, ravages, bone marrow suspension, vaginal flow, transcervical lavage, brain fluid, ascites, milk, secretions of the respiratory, intestinal and genitourinary tracts, and leukophoresis samples. In another embodiment, the maternal sample is a mixture of two or more biological samples *e.g.* a biological sample can comprise two or more of a biological fluid sample, a tissue sample, and a cell culture sample. In some embodiments, the sample is a sample that is easily obtainable by non-invasive procedures *e.g.* blood, plasma, serum, sweat, tears, sputum, urine, sputum, ear flow, saliva and feces. In some embodiments, the biological sample is a peripheral blood sample, or the plasma and serum fractions. In other embodiments, the biological sample is a swab or smear, a biopsy specimen, or a cell culture. As disclosed above, the terms "blood," "plasma" and "serum" expressly encompass fractions or processed portions thereof. Similarly, where a sample is taken from a biopsy, swab, smear, *etc.*, the "sample" expressly encompasses a processed fraction or portion derived from the biopsy, swab, smear, *etc.*

Samples can also be obtained from in vitro cultured tissues, cells, or other polynucleotide-containing sources. The cultured samples can be taken from sources including, but not limited to, cultures (e.g., tissue or cells) maintained in different media and conditions (e.g., pH, pressure, or temperature), cultures (e.g., tissue or cells) maintained for different periods of length, cultures (e.g., tissue or cells) treated with different factors or reagents (e.g., a drug candidate, or a modulator), or cultures of different types of tissue or cells.

Methods of isolating nucleic acids from biological sources are well known and will differ depending upon the nature of the source. One of skill in the art can readily isolate nucleic acid from a source as needed for the method described herein. In some instances, it can be advantageous to fragment the

nucleic acid molecules in the nucleic acid sample. Fragmentation can be random, or it can be specific, as achieved, for example, using restriction endonuclease digestion. Methods for random fragmentation are well known in the art, and include, for example, limited DNAse digestion, alkali treatment and physical shearing. In one embodiment, sample nucleic acids are obtained from as cfDNA, which is not subjected to fragmentation. In other embodiments, the sample nucleic acids are obtained as genomic DNA, which is subjected to fragmentation into fragments of approximately 500 or more base pairs, and to which NGS methods can be readily applied.

Determination of CNV for prenatal diagnoses

Cell-free fetal DNA and RNA circulating in maternal blood can be used for the early non-invasive prenatal diagnosis (NIPD) of an increasing number of genetic conditions, both for pregnancy management and to aid reproductive decision-making. The presence of cell-free DNA circulating in the bloodstream has been known for over 50 years. More recently, presence of small amounts of circulating fetal DNA was discovered in the maternal bloodstream during pregnancy (Lo *et al.,* Lancet 350:485-487 [1997]). Thought to originate from dying placental cells, cell-free fetal DNA (cfDNA) has been shown to consists of short fragments typically fewer than 200 bp in length Chan *et al.,* Clin Chem 50:88-92 [2004]), which can be discerned as early as 4 weeks gestation (Illanes *et al.,* Early Human Dev 83:563-566 [2007]), and known to be cleared from the maternal circulation within hours of delivery (Lo *et al.,* Am J Hum Genet 64:218-224 [1999]). In addition to cfDNA, fragments of cell-free fetal RNA (cfRNA) can also be discerned in the maternal bloodstream, originating from genes that are transcribed in the fetus or placenta. The extraction and subsequent analysis of these fetal genetic elements from a maternal blood sample offers novel opportunities for NIPD.

The present method is a polymorphism-independent method that for use in NIPD and that does not require that the fetal cfDNA be distinguished from the maternal cfDNA to enable the determination of a fetal aneuploidy. In some embodiments, the aneuploidy is a complete chromosomal trisomy or monosomy, or a partial trisomy or monosomy. Partial aneuploidies are caused by loss or gain of part of a chromosome, and encompass chromosomal imbalances resulting from unbalanced translocations, unbalanced inversions, deletions and insertions. By far, the most common known aneuploidy compatible with life is trisomy 21 *i.e.* Down Syndrome (DS), which is caused by the presence of part or all of chromosome 21. Rarely, DS can be cause by an inherited or sporadic defect whereby an extra copy of all or part of chromosome 21 becomes attached to another chromosome (usually chromosome 14) to form a single aberrant chromosome. DS is associated with intellectual impairment, severe learning difficulties and excess mortality caused by long-term health problems such as heart disease. Other aneuploidies with known clinical significance include Edward syndrome (trisomy 18) and Patau Syndrome (trisomy 13), which are frequently fatal within the first few months of life. Abnormalities associated with the number of sex chromosomes are also known and include monosomy X *e.g.* Turner syndrome (XO), and triple X syndrome (XXX) in female births

and Kleinefelter syndrome (XXY) and XYY syndrome in male births, which are all associated with various phenotypes including sterility and reduction in intellectual skills. The method of the invention can be used to diagnose these and other chromosomal abnormalities prenatally.

According to some embodiments of the present invention the trisomy determined by the present invention include without limitation trisomy 21 (T21; Down Syndrome), trisomy 18 (T18; Edward's Syndrome), trisomy 16 (T16), trisomy 22 (T22; Cat Eye Syndrome), trisomy 15 (T15; Prader Willi Syndrome), trisomy 13 (T13; Patau Syndrome), trisomy 8 (T8; Warkany Syndrome) and the XXY (Kleinefelter Syndrome), XYY, or XXX trisomies. It will be appreciated that various other complete trisomies and partial trisomies can be determined in fetal cfDNA according to the teachings of the present invention. Examples of partial trisomies include, but are not limited to, partial trisomy 1q32-44, trisomy 9 p with trisomy, trisomy 4 mosaicism, trisomy 17p, partial trisomy 4q26-qter, trisomy 9, partial 2p trisomy, partial trisomy 1q, and/or partial trisomy 6p/monosomy 6q.

The method of the present invention can be also used to determine chromosomal monosomy X, and partial monosomies such as, monosomy 13, monosomy 15, monosomy 16, monosomy 21, and monosomy 22, which are known to be involved in pregnancy miscarriage. Partial monosomy of chromosomes typically involved in complete aneuploidy can also be determined by the method of the invention. Monosomy 18p is a rare chromosomal disorder in which all or part of the short arm (p) of chromosome 18 is deleted (monosomic). The disorder is typically characterized by short stature, variable degrees of mental retardation, speech delays, malformations of the skull and facial (craniofacial) region, and/or additional physical abnormalities. Associated craniofacial defects may vary greatly in range and severity from case to case. Conditions caused by changes in the structure or number of copies of chromosome 15 include Angelman Syndrome and Prader-Willi Syndrome, which involve a loss of gene activity in the same part of chromosome 15, the 15q11-q13 region. It will be appreciated that several translocations and microdeletions can be asymptomatic in the carrier parent, yet can cause a major genetic disease in the offspring. For example, a healthy mother who carries the 15q11-q13 microdeletion can give birth to a child with Angelman syndrome, a severe neurodegenerative disorder. Thus, the present invention can be used to identify such a partial deletion and other deletions in the fetus. Partial monosomy 13q is a rare chromosomal disorder that results when a piece of the long arm (q) of chromosome 13 is missing (monosomic). Infants born with partial monosomy 13q may exhibit low birth weight, malformations of the head and face (craniofacial region), skeletal abnormalities (especially of the hands and feet), and other physical abnormalities. Mental retardation is characteristic of this condition. The mortality rate during infancy is high among individuals born with this disorder. Almost all cases of partial monosomy 13q occur randomly for no apparent reason (sporadic). 22q11.2 deletion syndrome, also known as DiGeorge syndrome, is a syndrome caused by the deletion of a small piece of chromosome 22. The deletion (22 q11.2) occurs near the middle of the chromosome on the long arm of one of the pair of chromosome. The features

of this syndrome vary widely, even among members of the same family, and affect many parts of the body. Characteristic signs and symptoms may include birth defects such as congenital heart disease, defects in the palate, most commonly related to neuromuscular problems with closure (velo-pharyngeal insufficiency), learning disabilities, mild differences in facial features, and recurrent infections. Microdeletions in chromosomal region 22q11.2 are associated with a 20 to 30-fold increased risk of schizophrenia. In one embodiment, the method of the invention is used to determine partial monosomies including but not limited to monosomy 18p, partial monosomy of chromosome 15 (15q11-q13), partial monosomy 13q, and partial monosomy of chromosome 22 can also be determined using the method.

The method of the invention can be also used to determine any aneuploidy if one of the parents is a known carrier of such abnormality. These include, but not limited to, mosaic for a small supernumerary marker chromosome (SMC); t(11;14)(p15;p13) translocation; unbalanced translocation t(8;11)(p23.2;p15.5); 11q23 microdeletion; Smith-Magenis syndrome 17p11.2 deletion; 22q13.3 deletion; Xp22.3 microdeletion; 10p14 deletion; 20p microdeletion, DiGeorge syndrome [del(22)(q11.2q11.23)], Williams syndrome (7q11.23 and 7q36 deletions); 1p36 deletion; 2p microdeletion; neurofibromatosis type 1 (17q11.2 microdeletion), Yq deletion ; Wolf-Hirschhorn syndrome (WHS, 4p16.3 microdeletion); 1p36.2 microdeletion; 11q14 deletion; 19q13.2 microdeletion; Rubinstein-Taybi (16 p13.3 microdeletion); 7p21 microdeletion; Miller-Dieker syndrome (17p13.3), 17p11.2 deletion; and 2q37 microdeletion.

Determination of Complete Fetal Chromosomal Aneuploidies

In one embodiment, the present invention provides a method for determining the presence or absence of any one or more different complete fetal chromosomal aneuploidies in a maternal test sample comprising fetal and maternal nucleic acid molecules. Preferably, the method determines the presence or absence of any four or more different complete chromosomal aneuploidies. The steps of the method comprise (a) obtaining sequence information for the fetal and maternal nucleic acids in the maternal test sample; and (b) using the sequence information to identify a number of sequence tags for each of any one or more chromosomes of interest selected from chromosomes 1-22, X and Y and to identify a number of sequence tags for a normalizing chromosome sequence for each of the any one or more chromosomes of interest. The normalizing chromosome sequence can be a single chromosome, or it can be a group of chromosomes selected from chromosomes 1-22, X, and Y. The method further uses in step (c) the number of sequence tags identified for each of the any one or more chromosomes of interest and the number of sequence tags identified for each normalizing chromosome sequence to calculate a single chromosome dose for each of the any one or more chromosomes of interest; and (d) compares each of the single chromosome doses for each of the any one or more chromosomes of interest to a threshold value for each of the one or more chromosomes

of interest, thereby determining the presence or absence of any one or more complete different fetal chromosomal aneuploidies in the maternal test sample.

In some embodiments, step (c) comprises calculating a single chromosome dose for each chromosomes of interest as the ratio of the number of sequence tags identified for each of the chromosomes of interest and the number of sequence tags identified for the normalizing chromosome for each of the chromosomes of interest.

In other embodiments, step (c) comprises calculating a single chromosome dose for each of the chromosomes of interest as the ratio of the number of sequence tags identified for each of the chromosomes of interest and the number of sequence tags identified for the normalizing chromosome for each of the chromosomes of interest. In other embodiments, step (c) comprises calculating a sequence tag ratio for a chromosome of interest by relating the number of sequence tags obtained for the chromosome of interest to the length of the chromosome of interest, and relating the number of tags for the corresponding normalizing chromosome sequence for the chromosome of interest to the length of the normalizing chromosome sequence, and calculating a chromosome dose for the chromosome of interest as a ratio of the sequence tags density of the chromosome of interest and the sequence tag density for the normalizing sequence. The calculation is repeated for each of all chromosomes of interest. Steps (a)-(d) can be repeated for test samples from different maternal subjects.

An example of the embodiment whereby four or more complete fetal chromosomal aneuploidies are determined in a maternal test sample comprising a mixture of fetal and maternal cell-free DNA molecules, comprises: (a) sequencing at least a portion of cell-free DNA molecules to obtain sequence information for the fetal and maternal cell-free DNA molecules in the test sample; (b) using the sequence information to identify a number of sequence tags for each of any twenty or more chromosomes of interest selected from chromosomes 1-22, X, and Y and to identify a number of sequence tags for a normalizing chromosome for each of the twenty or more chromosomes of interest; (c) using the number of sequence tags identified for each of the twenty or more chromosomes of interest and the number of sequence tags identified for each the normalizing chromosome to calculate a single chromosome dose for each of the twenty or more chromosomes of interest; and (d) comparing each of the single chromosome doses for each of the twenty or more chromosomes of interest to a threshold value for each of the twenty or more chromosomes of interest, and thereby determining the presence or absence of any twenty or more different complete fetal chromosomal aneuploidies in the test sample.

In another embodiment, the method for determining the presence or absence of any one or more different complete fetal chromosomal aneuploidies in a maternal test sample as described above uses a normalizing segment sequence for determining the dose of the chromosome of interest. In this instance, the method comprises (a)obtaining sequence information for said fetal and maternal nucleic

acids in said sample; (b) using said sequence information to identify a number of sequence tags for each of any one or more chromosomes of interest selected from chromosomes 1-22, X and Y and to identify a number of sequence tags for a normalizing segment sequence for each of said any one or more chromosomes of interest. The normalizing segment sequence can be a single segment of a chromosome or it can be a group of segments form one or more different chromosomes. The method further uses in step (c) the number of sequence tags identified for each of said any one or more chromosomes of interest and said number of sequence tags identified for said normalizing segment sequence to calculate a single chromosome dose for each of said any one or more chromosomes of interest; and (d) comparing each of said single chromosome doses for each of said any one or more chromosomes of interest to a threshold value for each of said one or more chromosomes of interest, and thereby determining the presence or absence of one or more different complete fetal chromosomal aneuploidies in said sample.

In some embodiments, step (c) comprises calculating a single chromosome dose for each of said chromosomes of interest as the ratio of the number of sequence tags identified for each of said chromosomes of interest and the number of sequence tags identified for said normalizing segment sequence for each of said chromosomes of interest.

In other embodiments, step (c) comprises calculating a sequence tag ratio for a chromosome of interest by relating the number of sequence tags obtained for the chromosome of interest to the length of the chromosome of interest, and relating the number of tags for the corresponding normalizing segment sequence for the chromosome of interest to the length of the normalizing segment sequence, and calculating a chromosome dose for the chromosome of interest as a ratio of the sequence tags density of the chromosome of interest and the sequence tag density for the normalizing segment sequence. The calculation is repeated for each of all chromosomes of interest. Steps (a)-(d) can be repeated for test samples from different maternal subjects.

A means for comparing chromosome doses of different sample sets is provided by determining a normalized chromosome value (NCV), which relates the chromosome dose in a test sample to the mean of the of the corresponding chromosome dose in a set of qualified samples. The NCV is calculated as:

$$NCV_{ij} = \frac{x_{ij} - \hat{\mu}_j}{\hat{\sigma}_j}$$

where $\hat{\mu}_j$ and $\hat{\sigma}_j$ are the estimated mean and standard deviation, respectively, for the j-th chromosome dose in a set of qualified samples, and x_{ij} is the observed j-th chromosome dose for test sample i.

In some embodiments, the presence or absence of at least one complete fetal chromosomal aneuploidy is determined. In other embodiments, the presence or absence of at least two, at least three, at least four, at least five, at least six, at least seven, at least eight, at least nine, at least ten, at least eleven, at least twelve, at least thirteen, at least fourteen, at least fifteen, at least sixteen, at least

seventeen, at least eighteen, at least nineteen, at least twenty, at least twenty-one, at least twenty-two, at least twenty-three, or twenty-four complete fetal chromosomal aneuploidies are determined in a sample, wherein twenty-two of the complete fetal chromosomal aneuploidies correspond to complete chromosomal aneuploidies of any one or more of the autosomes; the twenty-third and twenty fourth chromosomal aneuploidy correspond to a complete fetal chromosomal aneuploidy of chromosomes X and Y. As aneuploidies of sex chromosomes can comprise tetrasomies, pentasomies and other polysomies, the number of different complete chromosomal aneuploidies that can be determined according to the present method may be at least 24, at least 25, at least 26, at least 27, at least 28, at least 29, or at least 30 complete chromosomal aneuploidies. Thus, the number of different complete fetal chromosomal aneuploidies that are determined is related to the number of chromosomes of interest that are selected for analysis.

In one embodiment, determining the presence or absence of any one or more different complete fetal chromosomal aneuploidies in a maternal test sample as described above uses a normalizing segment sequence for one chromosome of interest, which is selected from chromosomes 1-22, X, and Y. In other embodiments, two or more chromosomes of interest are selected from any two or more of chromosomes 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16,17, 18, 19, 20, 21, 22, X, or Y. In one embodiment, any one or more chromosomes of interest are selected from chromosomes 1-22, X, and Y comprise at least twenty chromosomes selected from chromosomes 1-22, X, and Y, and wherein the presence or absence of at least twenty different complete fetal chromosomal aneuploidies is determined. In other embodiments, any one or more chromosomes of interest selected from chromosomes 1-22, X, and Y is all of chromosomes 1-22, X, and Y, and wherein the presence or absence of complete fetal chromosomal aneuploidies of all of chromosomes 1-22, X, and Y is determined. Complete different fetal chromosomal aneuploidies that can be determined include complete chromosomal trisomies, complete chromosomal monosomies and complete chromosomal polysomies. Examples of complete fetal chromosomal aneuploidies include without limitation trisomies of any one or more of the autosomes *e.g.* trisomy 2, trisomy 8, trisomy 9, trisomy 21, trisomy 13, trisomy 16, trisomy 18, trisomy 22; trisomies of the sex chromosomes *e.g.* 47,XXY, 47 XXX, and 47 XYY; tetrasomies of sex chromosomes *e.g.* 48,XXYY, 48,XXXY, 48XXXX, and 48,XYYY; pentasomies of sex chromosomes *e.g.* 49,XXXYY 49,XXXXY, 49,XXXXX, 49,XYYYY; and monosomy X. Other complete fetal chromosomal aneuploidies that can be determined according to the present method are described below.

Determination of Partial Fetal Chromosomal Aneuploidies

In another embodiment, the invention provides a method for determining the presence or absence of any one or more different partial fetal chromosomal aneuploidies in a maternal test sample comprising fetal and maternal nucleic acid molecules. The steps of the method comprise (a) obtaining sequence information for the fetal and maternal nucleic acids in said sample; and (b) using the

sequence information to identify a number of sequence tags for each of any one or more segments of any one or more chromosomes of interest selected from chromosomes 1-22, X, and Y and to identify a number of sequence tags for a normalizing segment sequence for each of said any one or more segments of any one or more chromosomes of interest. The normalizing segment sequence can be a single segment of a chromosome or it can be a group of segments form one or more different chromosomes. The method further uses in step (c) the number of sequence tags identified for each of any one or more segments of any one or more chromosomes of interest and the number of sequence tags identified for the normalizing segment sequence to calculate a single segment dose for each of any one or more segments of any one or more chromosome of interest; and (d) comparing each of the single chromosome doses for each of any one or more segments of any one or more chromosomes of interest to a threshold value for each of said any one or more chromosomal segments of any one or more chromosome of interest, and thereby determining the presence or absence of one or more different partial fetal chromosomal aneuploidies in said sample.

In some embodiments, step (c) comprises calculating a single segment dose for each of any one or more segments of any one or more chromosomes of interest as the ratio of the number of sequence tags identified for each of any one or more segments of any one or more chromosomes of interest and the number of sequence tags identified for the normalizing segment sequence for each of any one or more segments of any one or more chromosomes of interest.

In other embodiments, step (c) comprises calculating a sequence tag ratio for a segment of interest by relating the number of sequence tags obtained for the segment of interest to the length of the segment of interest, and relating the number of tags for the corresponding normalizing segment sequence for the segment of interest to the length of the normalizing segment sequence, and calculating a segment dose for the segment of interest as a ratio of the sequence tags density of the segment of interest and the sequence tag density for the normalizing segment sequence. The calculation is repeated for each of all chromosomes of interest. Steps (a)-(d) can be repeated for test samples from different maternal subjects.

A means for comparing segment doses of different sample sets is provided by determining a normalized segment value (NSV), which relates the segment dose in a test sample to the mean of the of the corresponding segment dose in a set of qualified samples. The NSV is calculated as:

$$NSV_{ij} = \frac{x_{ij} - \hat{\mu}_j}{\hat{\sigma}_j}$$

where $\hat{\mu}_j$ and $\hat{\sigma}_j$ are the estimated mean and standard deviation, respectively, for the *j*-th segment dose in a set of qualified samples, and x_{ij} is the observed *j*-th segment dose for test sample *i*.

In some embodiments, the presence or absence of one partial fetal chromosomal aneuploidy is determined. In other embodiments, the presence or absence of two, three, four, five, six, seven, eight,

nine, ten, fifteen, twenty, twenty-five, or more partial fetal chromosomal aneuplodies are determined in a sample. In one embodiment, one segment of interest selected from any one of chromosomes 1-22, X, and Y is selected from chromosomes 1-22, X, and Y. In another embodiment, two or more segments of interest selected from chromosomes 1-22, X, and Y are selected from any two or more of chromosomes 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16,17, 18, 19, 20, 21, 22, X, or Y. In one embodiment, any one or more segments of interest are selected from chromosomes 1-22, X, and Y comprise at least one, five, ten, 15, 20, 25 or more segments selected from chromosomes 1-22, X, and Y, and wherein the presence or absence of at least one, five, ten, 15, 20, 25 different partial fetal chromosomal aneuploidies is determined. Different partial fetal chromosomal aneuploidies that can be determined include fetal chromosomal aneuploidies include partial duplications, partial multiplications, partial insertions and partial deletions. Examples of partial fetal chromosomal aneuploidies include partial monosomies and partial trisomies of autosomes. Partial monosomies of autosomes include partial monosomy of chromosome 1, partial monosomy of chromosome 4, partial monosomy of chromosome 5, partial monosomy of chromosome 7, partial monosomy of chromosome 11, partial monosomy of chromosome 15, partial monosomy of chromosome 17, partial monosomy of chromosome 18, and partial monosomy of chromosome 22. Other partial fetal chromosomal aneuploidies that can be determined according to the present method are described below.

In any one of the embodiments described above, the test sample is a maternal sample selected from blood, plasma, serum, urine and saliva samples. In some embodiments, the maternal test sample is a plasma sample. The nucleic acid molecules of the maternal sample are a mixture of fetal and maternal cell-free DNA molecules. Sequencing of the nucleic acids can be performed using next generation sequencing (NGS) as described elsewhere herein. In some embodiments, sequencing is massively parallel sequencing using sequencing-by-synthesis with reversible dye terminators. In other embodiments, sequencing is sequencing-by-ligation. In yet other embodiments, sequencing is single molecule sequencing. Optionally, an amplification step is performed prior to sequencing.

Determination of CNV of clinical disorders

In addition to the early determination of birth defects, the methods described herein can be applied to the determination of any abnormality in the representation of genetic sequences within the genome.

It has been shown that blood plasma and serum DNA from cancer patients contains measurable quantities of tumor DNA, which can be recovered and used as surrogate source of tumor DNA, and tumors are characterized by aneuploidy, or inappropriate numbers of gene sequences or even entire chromosomes. The determination of a difference in the amount of a given sequence *i.e.* a sequence of interest, in a sample from an individual can thus be used in the diagnosis of a medical condition. In some embodiments, the method can be used to determine the presence or absence of a

chromosomal aneuploidy in a patient suspected or known to be suffering from cancer. The method can also be applied to determining the presence or absence of the status of a disease; to determining the presence or absence of nucleic acids of a pathogen *e.g.* virus; to determining chromosomal abnormalities associated with graft versus host disease (GVHD), and to determining the contribution of individuals in forensic analyses.

Embodiments of the invention provide for a method to assess copy number variation of a sequence of interest *e.g.* a clinically-relevant sequence, in a test sample that comprises a mixture of nucleic acids derived from two different genomes, and which are known or are suspected to differ in the amount of one or more sequence of interest. The mixture of nucleic acids is derived from two or more types of cells. In one embodiment, the mixture of nucleic acids is derived from normal and cancerous cells derived from a subject suffering from a medical condition *e.g.* cancer.

The development of cancer is often accompanied by an alteration in number of whole chromosomes *i.e.* complete chromosomal aneuploidy, and/or an alteration in the number of segments of chromosomes *i.e.* partial aneuploidy, caused by a process known as chromosome instability (CIN) (Thoma *et al.,* Swiss Med Weekly 2011:141:w13170). It is believed that many solid tumors, such as breast cancer, progress from initiation to metastasis through the accumulation of several genetic aberrations. [Sato *et al.,* Cancer Res., 50: 7184-7189 [1990]; Jongsma *et al.,* J Clin Pathol: Mol Path 55:305-309 [2002])]. Such genetic aberrations, as they accumulate, may confer proliferative advantages, genetic instability and the attendant ability to evolve drug resistance rapidly, and enhanced angiogenesis, proteolysis and metastasis. The genetic aberrations may affect either recessive "tumor suppressor genes" or dominantly acting oncogenes. Deletions and recombination leading to loss of heterozygosity (LOH) are believed to play a major role in tumor progression by uncovering mutated tumor suppressor alleles.

cfDNA has been found in the circulation of patients diagnosed with malignancies including but not limited to lung cancer (Pathak *et al.* Clin Chem 52:1833-1842 [2006]), prostate cancer (Schwartzenbach *et al.* Clin Cancer Res 15:1032-8 [2009]), and breast cancer (Schwartzenbach *et al.* available online at breast-cancer-research.com/content/11/5/R71 [2009]). Identification of genomic instabilities associated with cancers that can be determined in the circulating cfDNA in cancer patients is a potential diagnostic and prognostic tool. In one embodiment, the method of the invention assesses CNV of a sequence of interest in a sample comprising a mixture of nucleic acids derived from a subject that is suspected or is known to have cancer *e.g.* carcinoma, sarcoma, lymphoma, leukemia, germ cell tumors and blastoma. In one embodiment, the sample is a plasma sample derived (processes) from peripheral blood and that comprises a mixture of cfDNA derived from normal and cancerous cells. In another embodiment, the biological sample that is needed to determine whether a CNV is present is derived from a mixture of cancerous and non-cancerous cells from other biological fluids including but not limited to serum, sweat, tears, sputum, urine, sputum, ear flow, lymph, saliva,

cerebrospinal fluid, ravages, bone marrow suspension, vaginal flow, transcervical lavage, brain fluid, ascites, milk, secretions of the respiratory, intestinal and genitourinary tracts, and leukophoresis samples, or in tissue biopsies, swabs, or smears. In other embodiments, the biological sample is a stool (fecal) sample.

The sequence of interest is a nucleic acid sequence that is known or is suspected to play a role in the development and/or progression of the cancer. Examples of a sequence of interest include nucleic acids sequences *i.e.* complete chromosomes and/or segments of chromosomes, that are amplified or deleted in cancerous cells as described in the following.

In one embodiment, the present method can be used to determine the presence or absence of a chromosomal amplification. In some embodiments, the chromosomal amplification is the gain of one or more entire chromosomes. In other embodiments, the chromosomal amplification is the gain of one or more segments of a chromosome. In yet other embodiments, the chromosomal amplification is the gain of two or more segments of two or more chromosomes. The chromosomal amplification can involve the gain of one or more oncogenes.

Dominantly acting genes associated with human solid tumors typically exert their effect by overexpression or altered expression. Gene amplification is a common mechanism leading to upregulation of gene expression. Evidence from cytogenetic studies indicates that significant amplification occurs in over 50% of human breast cancers. Most notably, the amplification of the proto-oncogene human epidermal growth factor receptor 2 (HER2) located on chromosome 17 (17(17q21-q22)), results in overexpression of HER2 receptors on the cell surface leading to excessive and dysregulated signaling in breast cancer and other malignancies (Park *et al.,* Clinical Breast Cancer 8:392-401 [2008]). A variety of oncogenes have been found to be amplified in other human malignancies. Examples of the amplification of cellular oncogenes in human tumors include amplifications of: c-myc in promyelocytic leukemia cell line HL60, and in small-cell lung carcinoma cell lines, N-myc in primary neuroblastomas (stages III and IV), neuroblastoma cell lines, retinoblastoma cell line and primary tumors, and small-cell lung carcinoma lines and tumors, L-myc in small-cell lung carcinoma cell lines and tumors, c-myb in acute myeloid leukemia and in colon carcinoma cell lines, c-erbb in epidermoid carcinoma cell, and primary gliomas, c-K-ras-2 in primary carcinomas of lung, colon, bladder, and rectum, N-ras in mammary carcinoma cell line (Varmus H., Ann Rev Genetics 18: 553-612 (1984) [cited in Watson et al., Molecular Biology of the Gene (4th ed.; Benjamin/Cummings Publishing Co. 1987)].

In one embodiment, the present method can be used to determine the presence or absence of a chromosomal deletion. In some embodiments, the chromosomal deletion is the loss of one or more entire chromosomes. In other embodiments, the chromosomal deletion is the loss of one or more segments of a chromosome. In yet other embodiments, the chromosomal deletion is the loss of two or

more segments of two or more chromosomes. The chromosomal deletion can involve the loss of one or more tumor suppressor genes.

Chromosomal deletions involving tumor suppressor genes may play an important role in the development and progression of solid tumors. The retinoblastoma tumor suppressor gene (Rb-1), located in chromosome 13q14, is the most extensively characterized tumor suppressor gene. The Rb-1 gene product, a 105 kDa nuclear phosphoprotein, apparently plays an important role in cell cycle regulation (Howe et al., Proc Natl Acad Sci (USA) 87:5883-5887 [1990]). Altered or lost expression of the Rb protein is caused by inactivation of both gene alleles either through a point mutation or a chromosomal deletion. Rb-i gene alterations have been found to be present not only in retinoblastomas but also in other malignancies such as osteosarcomas, small cell lung cancer (Rygaard et al., Cancer Res 50: 5312-5317 [1990)]) and breast cancer. Restriction fragment length polymorphism (RFLP) studies have indicated that such tumor types have frequently lost heterozygosity at 13q suggesting that one of the Rb-1 gene alleles has been lost due to a gross chromosomal deletion (Bowcock et al., Am J Hum Genet, 46: 12 [1990]). Chromosome 1 abnormalities including duplications, deletions and unbalanced translocations involving chromosome 6 and other partner chromosomes indicate that regions of chromosome 1, in particular 1q21-1q32 and 1p11-13, might harbor oncogenes or tumor suppressor genes that are pathogenetically relevant to both chronic and advanced phases of myeloproliferative neoplasms (Caramazza *et al.,* Eur J Hematol84:191-200 [2010]). Myeloproliferative neoplasms are also associated with deletions of chromosome 5. Complete loss or interstitial deletions of chromosome 5 are the most common karyotypic abnormality in myelodysplastic syndromes (MDSs). Isolated del(5q)/5q- MDS patients have a more favorable prognosis than those with additional karyotypic defects, who tend to develop myeloproliferative neoplasms (MPNs) and acute myeloid leukemia. The frequency of unbalanced chromosome 5 deletions has led to the idea that 5q harbors one or more tumor-suppressor genes that have fundamental roles in the growth control of hematopoietic stem/progenitor cells (HSCs/HPCs). Cytogenetic mapping of commonly deleted regions (CDRs) centered on 5q31 and 5q32 identified candidate tumor-suppressor genes, including the ribosomal subunit RPS14, the transcription factor Egr1/Krox20 and the cytoskeletal remodeling protein, alpha-catenin (Eisenmann *et al.,* Oncogene 28:3429-3441 [2009]). Cytogenetic and allelotyping studies of fresh tumors and tumor cell lines have shown that allelic loss from several distinct regions on chromosome 3p, including 3p25, 3p21–22, 3p21.3, 3p12–13 and 3p14, are the earliest and most frequent genomic abnormalities involved in a wide spectrum of major epithelial cancers of lung, breast, kidney, head and neck, ovary, cervix, colon, pancreas, esophagus, bladder and other organs. Several tumor suppressor genes have been mapped to the chromosome 3p region, and are thought that interstitial deletions or promoter hypermethylation precede the loss of the 3p or the entire chromosome 3 in the development of carcinomas (Angeloni D., Briefings Functional Genomics 6:19-39 [2007]).

Newborns and children with Down syndrome (DS) often present with congenital transient leukemia and have an increased risk of acute myeloid leukemia and acute lymphoblastic leukemia. Chromosome 21, harboring about 300 genes, may be involved in numerous structural aberrations, e.g., translocations, deletions, and amplifications, in leukemias, lymphomas, and solid tumors. Moreover, genes located on chromosome 21 have been identified that play an important role in tumorigenesis. Somatic numerical as well as structural chromosome 21 aberrations are associated with leukemias, and specific genes including RUNX1, TMPRSS2, and TFF, which are located in 21q, play a role in tumorigenesis (Fonatsch C Gene Chromosomes Cancer 49:497-508 [2010]).

In one embodiment, the method provides a means to assess the association between gene amplification and the extent of tumor evolution. Correlation between amplification and/or deletion and stage or grade of a cancer may be prognostically important because such information may contribute to the definition of a genetically based tumor grade that would better predict the future course of disease with more advanced tumors having the worst prognosis. In addition, information about early amplification and/or deletion events may be useful in associating those events as predictors of subsequent disease progression. Gene amplification and deletions as identified by the method can be associated with other known parameters such as tumor grade, histology, Brd/Urd labeling index, hormonal status, nodal involvement, tumor size, survival duration and other tumor properties available from epidemiological and biostatistical studies. For example, tumor DNA to be tested by the method could include atypical hyperplasia, ductal carcinoma in situ, stage I-III cancer and metastatic lymph nodes in order to permit the identification of associations between amplifications and deletions and stage. The associations made may make possible effective therapeutic intervention. For example, consistently amplified regions may contain an overexpressed gene, the product of which may be able to be attacked therapeutically (for example, the growth factor receptor tyrosine kinase, $p185^{HER2}$).

The method can be used to identify amplification and/or deletion events that are associated with drug resistance by determining the copy number variation of nucleic acid sequences from primary cancers to those of cells that have metastasized to other sites. If gene amplification and/or deletion is a manifestation of karyotypic instability that allows rapid development of drug resistance, more amplification and/or deletion in primary tumors from chemoresistant patients than in tumors in chemosensitive patients would be expected. For example, if amplification of specific genes is responsible for the development of drug resistance, regions surrounding those genes would be expected to be amplified consistently in tumor cells from pleural effusions of chemoresistant patients but not in the primary tumors. Discovery of associations between gene amplification and/or deletion and the development of drug resistance may allow the identification of patients that will or will not benefit from adjuvant therapy.

In a manner similar to that described for determining the presence or absence of complete and/or partial fetal chromosomal aneuploidies in a maternal sample, the method of the invention can be used to determine the presence or absence of complete and/or partial chromosomal aneuploidies in any patient sample comprising nucleic acids *e.g.* DNA or cfDNA (including patient samples that are not maternal samples). The patient sample can be any biological sample type as described elsewhere herein. Preferably, the sample is obtained by non-invasive procedures. For example, the sample can be a blood sample, or the serum and plasma fractions thereof. Alternatively, the sample can be a urine sample or a fecal sample. In yet other embodiments, the sample is a tissue biopsy sample. In all cases, the sample comprises nucleic acids *e.g.* cfDNA or genomic DNA, which is purified, and sequenced using any of the NGS sequencing methods described previously.

Both complete and partial chromosomal aneuploidies associated with the formation, and progression of cancer can be determined according to the present method.

Determination of Complete Chromosomal Aneuploidies in Patient Samples

In one embodiment, the present invention provides a method for determining the presence or absence of any one or more different complete chromosomal aneuploidies in a patient test sample comprising nucleic acid molecules. In some embodiments, the method determines the presence or absence of any one or more different complete chromosomal aneuploidies. The steps of the method comprise (a) obtaining sequence information for the patient nucleic acids in the patient test sample; and (b) using the sequence information to identify a number of sequence tags for each of any one or more chromosomes of interest selected from chromosomes 1-22, X and Y and to identify a number of sequence tags for a normalizing chromosome sequence for each of the any one or more chromosomes of interest. The normalizing chromosome sequence can be a single chromosome, or it can be a group of chromosomes selected from chromosomes 1-22, X, and Y. The method further uses in step (c) the number of sequence tags identified for each of the any one or more chromosomes of interest and the number of sequence tags identified for each normalizing chromosome sequence to calculate a single chromosome dose for each of the any one or more chromosomes of interest; and (d) compares each of the single chromosome doses for each of the any one or more chromosomes of interest to a threshold value for each of the one or more chromosomes of interest, thereby determining the presence or absence of any one or more different complete patient chromosomal aneuploidies in the patient test sample.

In some embodiments, step (c) comprises calculating a single chromosome dose for each chromosomes of interest as the ratio of the number of sequence tags identified for each of the chromosomes of interest and the number of sequence tags identified for the normalizing chromosome for each of the chromosomes of interest.

In other embodiments, step (c) comprises calculating a single chromosome dose for each of the chromosomes of interest as the ratio of the number of sequence tags identified for each of the

chromosomes of interest and the number of sequence tags identified for the normalizing chromosome for each of the chromosomes of interest. In other embodiments, step (c) comprises calculating a sequence tag ratio for a chromosome of interest by relating the number of sequence tags obtained for the chromosome of interest to the length of the chromosome of interest, and relating the number of tags for the corresponding normalizing chromosome sequence for the chromosome of interest to the length of the normalizing chromosome sequence, and calculating a chromosome dose for the chromosome of interest as a ratio of the sequence tags density of the chromosome of interest and the sequence tag density for the normalizing sequence. The calculation is repeated for each of all chromosomes of interest. Steps (a)-(d) can be repeated for test samples from different patients.

An example of the embodiment whereby one or more complete chromosomal aneuploidies are determined in a cancer patient test sample comprising cell-free DNA molecules, comprises: (a) sequencing at least a portion of cell-free DNA molecules to obtain sequence information for the patient cell-free DNA molecules in the test sample; (b) using the sequence information to identify a number of sequence tags for each of any twenty or more chromosomes of interest selected from chromosomes 1-22, X, and Y and to identify a number of sequence tags for a normalizing chromosome for each of the twenty or more chromosomes of interest; (c) using the number of sequence tags identified for each of the twenty or more chromosomes of interest and the number of sequence tags identified for each the normalizing chromosome to calculate a single chromosome dose for each of the twenty or more chromosomes of interest; and (d) comparing each of the single chromosome doses for each of the twenty or more chromosomes of interest to a threshold value for each of the twenty or more chromosomes of interest, and thereby determining the presence or absence of any twenty or more different complete chromosomal aneuploidies in the patient test sample.

In another embodiment, the method for determining the presence or absence of any one or more different complete chromosomal aneuploidies in a patient test sample as described above uses a normalizing segment sequence for determining the dose of the chromosome of interest. In this instance, the method comprises (a) obtaining sequence information for the nucleic acids in the sample; (b) using the sequence information to identify a number of sequence tags for each of any one or more chromosomes of interest selected from chromosomes 1-22, X and Y and to identify a number of sequence tags for a normalizing segment sequence for each of any one or more chromosomes of interest. The normalizing segment sequence can be a single segment of a chromosome or it can be a group of segments form one or more different chromosomes. The method further uses in step (c) the number of sequence tags identified for each of said any one or more chromosomes of interest and said number of sequence tags identified for said normalizing segment sequence to calculate a single chromosome dose for each of said any one or more chromosomes of interest; and (d) comparing each of said single chromosome doses for each of said any one or more chromosomes of interest to a threshold value for each of said one or more chromosomes of interest, and thereby determining the

presence or absence of one or more different complete chromosomal aneuploidies in the patient sample.

In some embodiments, step (c) comprises calculating a single chromosome dose for each of said chromosomes of interest as the ratio of the number of sequence tags identified for each of said chromosomes of interest and the number of sequence tags identified for said normalizing segment sequence for each of said chromosomes of interest.

In other embodiments, step (c) comprises calculating a sequence tag ratio for a chromosome of interest by relating the number of sequence tags obtained for the chromosome of interest to the length of the chromosome of interest, and relating the number of tags for the corresponding normalizing segment sequence for the chromosome of interest to the length of the normalizing segment sequence, and calculating a chromosome dose for the chromosome of interest as a ratio of the sequence tags density of the chromosome of interest and the sequence tag density for the normalizing segment sequence. The calculation is repeated for each of all chromosomes of interest. Steps (a)-(d) can be repeated for test samples from different patients.

A means for comparing chromosome doses of different sample sets is provided by determining a normalized chromosome value (NCV), which relates the chromosome dose in a test sample to the mean of the of the corresponding chromosome dose in a set of qualified samples. The NCV is calculated as:

$$NCV_{ij} = \frac{x_{ij} - \hat{\mu}_j}{\hat{\sigma}_j}$$

where $\hat{\mu}_j$ and $\hat{\sigma}_j$ are the estimated mean and standard deviation, respectively, for the j-th chromosome dose in a set of qualified samples, and x_{ij} is the observed j-th chromosome dose for test sample i.

In some embodiments, the presence or absence of one complete chromosomal aneuploidy is determined. In other embodiments, the presence or absence of two, three, four, five, six, seven, eight, nine, ten, eleven, twelve, thirteen, fourteen, fifteen, sixteen, seventeen, eighteen, nineteen, twenty, twenty-one, twenty-two, twenty-three, or twenty four complete chromosomal aneuploidies are determined in a sample, wherein twenty-two of the complete chromosomal aneuploidies correspond to complete chromosomal aneuploidies of any one or more of the autosomes; the twenty-third and twenty fourth chromosomal aneuploidy correspond to a complete chromosomal aneuploidy of chromosomes X and Y. As aneuploidies can comprise trisomies, tetrasomies, pentasomies and other polysomies, and the number of complete chromosomal aneuploidies varies in different diseases and in different stages of the same disease, the number of complete chromosomal aneuploidies that are determined according to the present method are at least 24, at least 25, at least 26, at least 27, at least 28, at least 29, at least 30 complete, at least 40, at least 50, at least 60, at least 70, at least 80, at least 90, at least 100 or more chromosomal aneuploidies. Systematic karyotyping of tumors has revealed that the chromosome number in cancer cells is highly variable, ranging from hypodiploidy

(considerably fewer than 46 chromosomes) to tetraploidy and hypertetraploidy (up to 200 chromosomes) (Storchova and Kuffer J Cell Sci 121:3859-3866 [2008]). In some embodiments, the method comprises determining the presence or absence of up to 200 or more chromosomal aneuploidies in a sample form a patient suspected or known to be suffering from cancer *e.g.* colon cancer. The chromosomal aneuploidies include losses of one or more complete chromosomes (hypodiploidies), gains of complete chromosomes including trisomies, tetrasomies, pentasomies, and other polysomies. Gains and/or losses of segments of chromosomes can also be determined as described elsewhere herein. The method is applicable to determining the presence or absence of different aneuploidies in samples from patients suspected or known to be suffering from any cancer as described elsewhere herein.

In some embodiments, any one of chromosomes 1-22, X and Y, can be the chromosome of interest in determining the presence or absence of any one or more different complete chromosomal aneuploidies in a patient test sample as described above. In other embodiments, two or more chromosomes of interest are selected from any two or more of chromosomes 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16,17, 18, 19, 20, 21, 22, X, or Y. In one embodiment, any one or more chromosomes of interest are selected from chromosomes 1-22, X, and Y comprise at least twenty chromosomes selected from chromosomes 1-22, X, and Y, and wherein the presence or absence of at least twenty different complete chromosomal aneuploidies is determined. In other embodiments, any one or more chromosomes of interest selected from chromosomes 1-22, X, and Y is all of chromosomes 1-22, X, and Y, and wherein the presence or absence of complete chromosomal aneuploidies of all of chromosomes 1-22, X, and Y is determined. Complete different chromosomal aneuploidies that can be determined include complete chromosomal monosomies of any one or more of chromosomes 1-22, X and Y; complete chromosomal trisomies of any one or more of chromosomes 1-22, X and Y; complete chromosomal tetrasomies of any one or more of chromosomes 1-22, X and Y; complete chromosomal pentasomies of any one or more of chromosomes 1-22, X and Y; and other complete chromosomal polysomies of any one or more of chromosomes 1-22, X and Y.

Determination of Partial Chromosomal Aneuploidies in Patient Samples

In another embodiment, the invention provides a method for determining the presence or absence of any one or more different partial chromosomal aneuploidies in a patient test sample comprising nucleic acid molecules. The steps of the method comprise (a) obtaining sequence information for the patient nucleic acids in the sample; and (b) using the sequence information to identify a number of sequence tags for each of any one or more segments of any one or more chromosomes of interest selected from chromosomes 1-22, X, and Y and to identify a number of sequence tags for a normalizing segment sequence for each of any one or more segments of any one or more chromosomes of interest. The normalizing segment sequence can be a single segment of a chromosome or it can be a group of segments form one or more different chromosomes. The method

further uses in step (c) the number of sequence tags identified for each of any one or more segments of any one or more chromosomes of interest and the number of sequence tags identified for the normalizing segment sequence to calculate a single segment dose for each of any one or more segments of any one or more chromosome of interest; and (d) comparing each of the single chromosome doses for each of any one or more segments of any one or more chromosomes of interest to a threshold value for each of said any one or more chromosomal segments of any one or more chromosome of interest, and thereby determining the presence or absence of one or more different partial chromosomal aneuploidies in said sample.

In some embodiments, step (c) comprises calculating a single segment dose for each of any one or more segments of any one or more chromosomes of interest as the ratio of the number of sequence tags identified for each of any one or more segments of any one or more chromosomes of interest and the number of sequence tags identified for the normalizing segment sequence for each of any one or more segments of any one or more chromosomes of interest.

In other embodiments, step (c) comprises calculating a sequence tag ratio for a segment of interest by relating the number of sequence tags obtained for the segment of interest to the length of the segment of interest, and relating the number of tags for the corresponding normalizing segment sequence for the segment of interest to the length of the normalizing segment sequence, and calculating a segment dose for the segment of interest as a ratio of the sequence tags density of the segment of interest and the sequence tag density for the normalizing segment sequence. The calculation is repeated for each of all chromosomes of interest. Steps (a)-(d) can be repeated for test samples from different patients.

A means for comparing segment doses of different sample sets is provided by determining a normalized segment value (NSV), which relates the segment dose in a test sample to the mean of the of the corresponding segment dose in a set of qualified samples. The NSV is calculated as:

$$NSV_{ij} = \frac{x_{ij} - \hat{\mu}_j}{\hat{\sigma}_j}$$

where $\hat{\mu}_j$ and $\hat{\sigma}_j$ are the estimated mean and standard deviation, respectively, for the *j*-th segment dose in a set of qualified samples, and x_{ij} is the observed *j*-th segment dose for test sample *i*.

In some embodiments, the presence or absence of one partial chromosomal aneuploidy is determined. In other embodiments, the presence or absence of two, three, four, five, six, seven, eight, nine, ten, fifteen, twenty, twenty-five, or more partial chromosomal aneuploidies are determined in a sample. In one embodiment, one segment of interest selected from any one of chromosomes 1-22, X, and Y is selected from chromosomes 1-22, X, and Y. In another embodiment, two or more segments of interest selected from chromosomes 1-22, X, and Y are selected from any two or more of chromosomes 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16,17, 18, 19, 20, 21, 22, X, or Y. In one

embodiment, any one or more segments of interest are selected from chromosomes 1-22, X, and Y comprise at least one, five, ten, 15, 20, 25, 50, 75, 100 or more segments selected from chromosomes 1-22, X, and Y, and wherein the presence or absence of at least one, five, ten, 15, 20, 25, 50, 75, 100, or more different partial chromosomal aneuploidies is determined. Different partial chromosomal aneuploidies that can be determined include chromosomal aneuploidies include partial duplications, partial multiplications, partial insertions and partial deletions.

Samples that can be used for determining the presence or absence of a chromosomal aneuploidy (partial or complete) in a patient can be any of the biological samples described elsewhere herein. The type of sample or samples that can be used for the determination of aneuploidy in a patient will depend on the type of disease from which the patient is known or suspected to be suffering. For example, a stool sample can be chosen as a source of DNA to determine the presence or absence of aneuploidies associated with colorectal cancer. The method is also applicable to tissue samples as described herein. Preferably, the sample is a biological sample that is obtained by non-invasive means *e.g.* a plasma sample. As described elsewhere herein, sequencing of the nucleic acids in the patient sample can be performed using next generation sequencing (NGS) as described elsewhere herein. In some embodiments, sequencing is massively parallel sequencing using sequencing-by-synthesis with reversible dye terminators. In other embodiments, sequencing is sequencing-by-ligation. In yet other embodiments, sequencing is single molecule sequencing. Optionally, an amplification step is performed prior to sequencing.

In some embodiments, the presence or absence of an aneuploidy is determined in a patient suspected to be suffering from a cancer as described elsewhere herein *e.g.* lung, breast, kidney, head and neck, ovary, cervix, colon, pancreas, esophagus, bladder and other organs, and blood cancers. Blood cancers include cancers of the bone marrow, blood, and lymphatic system, which includes lymph nodes, lymphatic vessels, tonsils, thymus, spleen, and digestive tract lymphoid tissue. Leukemia and myeloma, which start in the bone marrow, and lymphoma, which starts in the lymphatic system, are the most common types of blood cancer.

The determination of the presence or absence of one or more chromosomal aneuploidies in a patient sample can be made without limitation to determine the predisposition of the patient to a particular cancer, to determine the presence or absence of a cancer as part of routine screen in patients known and not known to be predisposed to the cancer in question, to provide a prognosis for the disease, to assess the need for adjuvant therapy, and to determine the progress or regress of the diseases.

Apparatus and systems for determining CNV

Analysis of the sequencing data and the diagnosis derived therefrom are typically performed using various computer algorithms and programs. In one embodiment, the invention provides a computer program product for generating an output indicating the presence or absence of a fetal

aneuploidy in a test sample. The computer product comprises a computer readable medium having a computer executable logic recorded thereon for enabling a processor to diagnose a fetal aneuploidy comprising: a receiving procedure for receiving sequencing data from at least a portion of nucleic acid molecules from a maternal biological sample, wherein said sequencing data comprises a calculated chromosome; computer assisted logic for analyzing a fetal aneuploidy from said received data; and an output procedure for generating an output indicating the presence, absence or kind of said fetal aneuploidy.

The method of the invention can be performed using a computer-readable medium having stored thereon computer-readable instructions for carrying out a method for identifying any CNV *e.g.* chromosomal or partial aneuploidies. Thus, in one embodiment, the invention provides a computer-readable medium having stored thereon computer-readable instructions for carrying out a method for identifying complete and partial chromosomal aneuploidies *e.g.* fetal aneuploidies.

The method of the invention can also be performed using a computer processing system which is adapted or configured to perform a method for identifying any CNV *e.g.* chromosomal or partial aneuploidies. Thus, in one embodiment, the invention provides a computer processing system which is adapted or configured to perform a method as described herein. In one embodiment, the apparatus comprises a sequencing device adapted or configured for sequencing at least a portion of the nucleic acid molecules in a sample to obtain the type of sequence information described elsewhere herein.

The present invention is described in further detail in the following Examples which are not in any way intended to limit the scope of the invention as claimed. The attached Figures are meant to be considered as integral parts of the specification and description of the invention. The following examples are offered to illustrate, but not to limit the claimed invention.

EXPERIMENTAL

Example 1

Sample processing and DNA extraction

Peripheral blood samples were collected from pregnant women in their first or second trimester of pregnancy and who were deemed at risk for fetal aneuploidy. Informed consent was obtained from each participant prior to the blood draw. Blood was collected before amniocentesis or chorionic villus sampling. Karyotype analysis was performed using the chorionic villus or amniocentesis samples to confirm fetal karyotype.

Peripheral blood drawn from each subject was collected in ACD tubes. One tube of blood sample (approximately 6-9 mL/tube) was transferred into one 15-mL low speed centrifuge tube. Blood was centrifuged at 2640 rpm, 4ºC for 10 min using Beckman Allegra 6 R centrifuge and rotor model GA 3.8.

For cell-free plasma extraction, the upper plasma layer was transferred to a 15-ml high speed centrifuge tube and centrifuged at 16000 x g, 4ºC for 10 min using Beckman Coulter Avanti J-E centrifuge, and JA-14 rotor. The two centrifugation steps were performed within 72 h after blood collection. Cell-free plasma was stored at -80ºC and thawed only once before DNA extraction.

Cell-free DNA was extracted from cell-free plasma by using QIAamp DNA Blood Mini kit (Qiagen) according to the manufacturer's instructions. Five milliliters of buffer AL and 500 µl of Qiagen Protease were added to 4.5 ml – 5 ml of cell-free plasma. The volume was adjusted to 10ml with phosphate buffered saline (PBS), and the mixture was incubated at 56°C for 12 minutes. Multiple columns were used to separate the precipitated cfDNA from the solution by centrifugation at 8,000 RPM in a Beckman microcentrifuge. The columns were washed with AW1 and AW2 buffers, and the cfDNA was eluted with 55 µl of nuclease-free water. Approximately 3.5-7 ng of cfDNA was extracted from the plasma samples.

All sequencing libraries were prepared from approximately 2 ng of purified cfDNA that was extracted from maternal plasma. Library preparation was performed using reagents of the NEBNext™ DNA Sample Prep DNA Reagent Set 1 (Part No. E6000L; New England Biolabs, Ipswich, MA), for Illumina® as follows. Because cell-free plasma DNA is fragmented in nature, no further fragmentation by nebulization or sonication was done on the plasma DNA samples. The overhangs of approximately 2 ng purified cfDNA fragments contained in 40µl were converted into phosphorylated blunt ends according to the NEBNext® End Repair Module by incubating in a 1.5ml microfuge tube the cfDNA with 5µl 10X phosphorylation buffer, 2µl deoxynucleotide solution mix (10 mM each dNTP), 1µl of a 1:5 dilution of DNA Polymerase I, 1µl T4 DNA Polymerase and 1µl T4 Polynucleotide Kinase provided in the NEBNext™ DNA Sample Prep DNA Reagent Set 1 for 15 minutes at 20°C. The enzymes were then heat inactivated by incubating the reaction mixture at 75°C for 5 minutes. The mixture was cooled to 4°C, and dA tailing of the blunt-ended DNA was accomplished using 10µl of the dA-tailing master mix containing the Klenow fragment (3' to 5' exo minus) (NEBNext™ DNA Sample Prep DNA Reagent Set 1), and incubating for 15 minutes at 37°C. Subsequently, the Klenow fragment was heat inactivated by incubating the reaction mixture at 75°C for 5 minutes. Following the inactivation of the Klenow fragment, 1µl of a 1:5 dilution of Illumina Genomic Adaptor Oligo Mix (Part No. 1000521; Illumina Inc., Hayward, CA) was used to ligate the Illumina adaptors (Non-Index Y-Adaptors) to the dA-tailed DNA using 4µl of the T4 DNA ligase provided in the NEBNext™ DNA Sample Prep DNA Reagent Set 1, by incubating the reaction mixture for 15 minutes at 25°C. The mixture was cooled to 4°C, and the adaptor-ligated cfDNA was purified from unligated adaptors, adaptor dimers, and other reagents using magnetic beads provided in the Agencourt AMPure XP PCR purification system (Part No. A63881; Beckman Coulter Genomics, Danvers, MA). Eighteen cycles of PCR were performed to selectively enrich adaptor-ligated cfDNA using Phusion ® High-Fidelity Master Mix (Finnzymes, Woburn, MA) and Illumina's PCR primers

complementary to the adaptors (Part No. 1000537 and 1000537). The adaptor-ligated DNA was subjected to PCR (98°C for 30 seconds; 18 cycles of 98°C for 10 seconds, 65°C for 30 seconds, and 72°C for 30 seconds; final extension at 72°C for 5 minutes, and hold at 4°C) using Illumina Genomic PCR Primers (Part Nos. 100537 and 1000538) and the Phusion HF PCR Master Mix provided in the NEBNext™ DNA Sample Prep DNA Reagent Set 1, according to the manufacturer's instructions. The amplified product was purified using the Agencourt AMPure XP PCR purification system (Agencourt Bioscience Corporation, Beverly, MA) according to the manufacturer's instructions available at www.beckmangenomics.com/products/AMPureXPProtocol_000387v001.pdf. The purified amplified product was eluted in 40µl of Qiagen EB Buffer, and the concentration and size distribution of the amplified libraries was analyzed using the Agilent DNA 1000 Kit for the 2100 Bioanalyzer (Agilent technologies Inc., Santa Clara, CA).

The amplified DNA was sequenced using Illumina's Genome Analyzer II to obtain single-end reads of 36bp. Only about 30 bp of random sequence information are needed to identify a sequence as belonging to a specific human chromosome. Longer sequences can uniquely identify more particular targets. In the present case, a large number of 36 bp reads were obtained, covering approximately 10% of the genome. Upon completion of sequencing of the sample, the Illumina "Sequencer Control Software" transferred image and base call files to a Unix server running the Illumina "Genome Analyzer Pipeline" software version 1.51. The Illumina "Gerald" program was run to align sequences to the reference human genome that is derived from the hg18 genome provided by National Center for Biotechnology Information (NCBI36/hg18, available on the world wide web at http://genome.ucsc.edu/cgi-bin/hgGateway?org=Human&db=hg18&hgsid=166260105). The sequence data generated from the above procedure that uniquely aligned to the genome was read from Gerald output (export.txt files) by a program (c2c.pl) running on a computer running the Linnux operating system. Sequence alignments with base mis-matches were allowed and included in alignment counts only if they aligned uniquely to the genome. Sequence alignments with identical start and end coordinates (duplicates) were excluded.

Between about 5 and 15 million 36 bp tags with 2 or less mismatches were mapped uniquely to the human genome. All mapped tags were counted and included in the calculation of chromosome doses in both test and qualifying samples. Regions extending from base 0 to base 2×10^6, base 10×10^6 to base 13×10^6, and base 23×10^6 to the end of chromosome Y, were specifically excluded from the analysis because tags derived from either male or female fetuses map to these regions of the Y-chromosome.

It was noted that some variation in the total number of sequence tags mapped to individual chromosomes across samples sequenced in the same run (inter-chromosomal variation), but substantially greater variation was noted to occur among different sequencing runs (inter-sequencing run variation).

Example 2

Dose and variance for chromosomes 13, 18, 21, X, and Y

To examine the extent of inter-chromosomal and inter-sequencing variation in the number of mapped sequence tags for all chromosomes, plasma cfDNA obtained from peripheral blood of 48 volunteer pregnant subjects was extracted and sequenced as described in Example 1, and analyzed as follows.

The total number of sequence tags that were mapped to each chromosome (sequence tag density) was determined. Alternatively, the number of mapped sequence tags may be normalized to the length of the chromosome to generate a sequence tag density ratio. The normalization to chromosome length is not a required step, and can be performed solely to reduce the number of digits in a number to simplify it for human interpretation. Chromosome lengths that can be used to normalize the sequence tags counts can be the lengths provided on the world wide web at genome.ucsc.edu/goldenPath/stats.html#hg18.

The resulting sequence tag density for each chromosome was related to the sequence tag density of each of the remaining chromosomes to derive a qualified chromosome dose, which was calculated as the ratio of the sequence tag density for the chromosome of interest *e.g.* chromosome 21, and the sequence tag density of each of the remaining chromosomes *i.e.* chromosomes 1-20, 22 and X. Table 1 provides an example of the calculated qualified chromosome dose for chromosomes of interest 13, 18, 21, X, and Y, determined in one of the qualified samples. Chromosomes doses were determined for all chromosomes in all samples, and the average doses for chromosomes of interest 13, 18, 21, X and Y in the qualified samples are provided in Tables 2 and 3, and depicted in Figures 2-6. Figures 2-6 also depict the chromosome doses for the test samples. The chromosome doses for each of the chromosomes of interest in the qualified samples provides a measure of the variation in the total number of mapped sequence tags for each chromosome of interest relative to that of each of the remaining chromosomes. Thus, qualified chromosome doses can identify the chromosome or a group of chromosomes *i.e.* normalizing chromosome that has a variation among samples that is closest to the variation of the chromosome of interest, and that would serve as ideal sequences for normalizing values for further statistical evaluation. Figures 7 and 8 depict the calculated average chromosome doses determined in a population of qualified samples for chromosomes 13, 18, and 21, and chromosomes X and Y.

In some instances, the best normalizing chromosome may not have the least variation, but may have a distribution of qualified doses that best distinguishes a test sample or samples from the qualified samples *i.e.* the best normalizing chromosome may not have the lowest variation, but may have the greatest differentiability. Thus, differentiability accounts for the variation in chromosome dose and the distribution of the doses in the qualified samples.

Tables 2 and 3 provide the coefficient of variation as the measure of variability, and student t-test values as a measure of differentiability for chromosomes 18, 21, X and Y, wherein the smaller the T-test value, the greatest the differentiability. The differentiability for chromosome 13 was determined as the ratio of difference between the mean chromosome dose in the qualified samples and the dose for chromosome 13 in the only T13 test sample, and the standard deviation of mean of the qualified dose.

The qualified chromosome doses also serve as the basis for determining threshold values when identifying aneuploidies in test samples as described in the following.

TABLE 1

Qualified Chromosome Dose for Chromosomes 13, 18, 21, X and Y (n=1; sample #11342, 46 XY)

Chromosome	chr 21	chr 18	chr 13	chr X	chrY
chr1	0.149901	0.306798	0.341832	0.490969	0.003958
chr2	0.15413	0.315452	0.351475	0.504819	0.004069
chr3	0.193331	0.395685	0.44087	0.633214	0.005104
chr4	0.233056	0.476988	0.531457	0.763324	0.006153
chr5	0.219209	0.448649	0.499882	0.717973	0.005787
chr6	0.228548	0.467763	0.521179	0.748561	0.006034
chr7	0.245124	0.501688	0.558978	0.802851	0.006472
chr8	0.256279	0.524519	0.584416	0.839388	0.006766
chr9	0.309871	0.634203	0.706625	1.014915	0.008181
chr10	0.25122	0.514164	0.572879	0.822817	0.006633
chr11	0.257168	0.526338	0.586443	0.8423	0.00679
chr12	0.275192	0.563227	0.627544	0.901332	0.007265
chr13	0.438522	0.897509	1	1.436285	0.011578
chr14	0.405957	0.830858	0.925738	1.329624	0.010718
chr15	0.406855	0.832697	0.927786	1.332566	0.010742
chr16	0.376148	0.769849	0.857762	1.231991	0.009931
chr17	0.383027	0.783928	0.873448	1.254521	0.010112
chr18	0.488599	1	1.114194	1.600301	0.0129
chr19	0.535867	1.096742	1.221984	1.755118	0.014148
chr20	0.467308	0.956424	1.065642	1.530566	0.012338
chr21	1	2.046668	2.280386	3.275285	0.026401
chr22	0.756263	1.547819	1.724572	2.476977	0.019966

chrX	0.305317	0.624882	0.696241	1	0.008061
chrY	37.87675	77.52114	86.37362	124.0572	1

TABLE 2

Qualified Chromosome Dose, Variance and Differentiability for chromosomes 21, 18 and 13

	21 (n=35)				**18 (n=40)**			
	Avg	Stdev	CV	T Test	Avg	Stdev	CV	T Test
chr1	0.15335	0.001997	1.30	3.18E-10	0.31941	0.008384	2.62	0.001675
chr2	0.15267	0.001966	1.29	9.87E-07	0.31807	0.001756	0.55	4.39E-05
chr3	0.18936	0.004233	2.24	1.04E-05	0.39475	0.002406	0.61	3.39E-05
chr4	0.21998	0.010668	4.85	0.000501	0.45873	0.014292	3.12	0.001349
chr5	0.21383	0.005058	2.37	1.43E-05	0.44582	0.003288	0.74	3.09E-05
chr6	0.22435	0.005258	2.34	1.48E-05	0.46761	0.003481	0.74	2.32E-05
chr7	0.24348	0.002298	0.94	2.05E-07	0.50765	0.004669	0.92	9.07E-05
chr8	0.25269	0.003497	1.38	1.52E-06	0.52677	0.002046	0.39	4.89E-05
chr9	0.31276	0.003095	0.99	3.83E-09	0.65165	0.013851	2.13	0.000559
chr10	0.25618	0.003112	1.21	2.28E-10	0.53354	0.013431	2.52	0.002137
chr11	0.26075	0.00247	0.95	1.08E-09	0.54324	0.012859	2.37	0.000998
chr12	0.27563	0.002316	0.84	2.04E-07	0.57445	0.006495	1.13	0.000125
chr13	0.41828	0.016782	4.01	0.000123	0.87245	0.020942	2.40	0.000164
chr14	0.40671	0.002994	0.74	7.33E-08	0.84731	0.010864	1.28	0.000149
chr15	0.41861	0.007686	1.84	1.85E-10	0.87164	0.027373	3.14	0.003862
chr16	0.39977	0.018882	4.72	7.33E-06	0.83313	0.050781	6.10	0.075458
chr17	0.41394	0.02313	5.59	0.000248	0.86165	0.060048	6.97	0.088579
chr18	0.47236	0.016627	3.52	1.3E-07				
chr19	0.59435	0.05064	8.52	0.01494	1.23932	0.12315	9.94	0.231139
chr20	0.49464	0.021839	4.42	2.16E-06	1.03023	0.058995	5.73	0.061101
chr21					2.03419	0.08841	4.35	2.81E-05
chr22	0.84824	0.070613	8.32	0.02209	1.76258	0.169864	9.64	0.181808
chrX	0.27846	0.015546	5.58	0.000213	0.58691	0.026637	4.54	0.064883

TABLE 3

Qualified Chromosome Dose, Variance and Differentiability for chromosomes 13, X, and Y

	13 (n=47)				X (n=19)			
	Avg	**Stdev**	**CV**	**Diff**	**Avg**	**Stdev**	**CV**	**T Test**
chr1	0.36536	0.01775	4.86	1.904	0.56717	0.025988	4.58	0.001013
chr2	0.36400	0.009817	2.70	2.704	0.56753	0.014871	2.62	
chr3	0.45168	0.007809	1.73	3.592	0.70524	0.011932	1.69	
chr4	0.52541	0.005264	1.00	3.083	0.82491	0.010537	1.28	
chr5	0.51010	0.007922	1.55	3.944	0.79690	0.012227	1.53	1.29E-11
chr6	0.53516	0.008575	1.60	3.758	0.83594	0.013719	1.64	2.79E-11
chr7	0.58081	0.017692	3.05	2.445	0.90507	0.026437	2.92	7.41E-07
chr8	0.60261	0.015434	2.56	2.917	0.93990	0.022506	2.39	2.11E-08
chr9	0.74559	0.032065	4.30	2.102	1.15822	0.047092	4.07	0.000228
chr10	0.61018	0.029139	4.78	2.060	0.94713	0.042866	4.53	0.000964
chr11	0.62133	0.028323	4.56	2.081	0.96544	0.041782	4.33	0.000419
chr12	0.65712	0.021853	3.33	2.380	1.02296	0.032276	3.16	3.95E-06
chr13					1.56771	0.014258	0.91	2.47E-15
chr14	0.96966	0.034017	3.51	2.233	1.50951	0.05009	3.32	8.24E-06
chr15	0.99673	0.053512	5.37	1.888	1.54618	0.077547	5.02	0.002925
chr16	0.95169	0.080007	8.41	1.613	1.46673	0.117073	7.98	0.114232
chr17	0.98547	0.091918	9.33	1.484	1.51571	0.132775	8.76	0.188271
chr18	1.13124	0.040032	3.54	2.312	1.74146	0.072447	4.16	0.001674
chr19	1.41624	0.174476	12.32	1.306	2.16586	0.252888	11.68	0.460752
chr20	1.17705	0.094807	8.05	1.695	1.81576	0.137494	7.57	0.08801
chr21	2.33660	0.131317	5.62	1.927	3.63243	0.235392	6.48	0.00675
chr22	2.01678	0.243883	12.09	1.364	3.08943	0.34981	11.32	0.409449
chrX	0.66679	0.028788	4.32	1.114				
chr2-6	0.46751	0.006762	1.45	4.066				
chr3-6	0.50332	0.005161	1.03	5.260				
chr_tot					1.13209	0.038485	3.40	2.7E-05
	Y (n=26)							

	Avg	Stdev	CV	T Test	
Chr 1-22, X	0.00734	0.002611	30.81	1.8E-12	

Examples of diagnoses of T21, T13, T18 and a case of Turner syndrome obtained using the normalizing chromosomes, chromosome doses and differentiability for each of the chromosomes of interest are described in Example 3.

EXAMPLE 3

Diagnosis of Fetal Aneuploidy Using Normalizing Chromosomes

To apply the use of chromosome doses for assessing aneuploidy in a biological test sample, maternal blood test samples were obtained from pregnant volunteers and cfDNA was prepared, sequenced and analyzed as described in Examples 1 and 2.

Trisomy 21

Table 4 provides the calculated dose for chromosome 21 in an exemplary test sample (#11403). The calculated threshold for the positive diagnosis of T21 aneuploidy was set at > 2 standard deviations from the mean of the qualified (normal) samples. A diagnosis for T21 was given based on the chromosome dose in the test sample being greater than the set threshold. Chromosomes 14 and 15 were used as normalizing chromosomes in separate calculations to show that either a chromosome having the lowest variability *e.g.* chromosome 14, or a chromosome having the greatest differentiability *e.g.* chromosome 15, can be used to identify the aneuploidy. Thirteen T21 samples were identified using the calculated chromosome doses, and the aneuploidy samples were confirmed to be T21 by karyotype.

TABLE 4

Chromosome Dose for a T21 aneuploidy (sample #11403, 47 XY +21)

Chromosome	Sequence Tag Density	Chromosome Dose for Chr 21	Threshold
Chr21	333,660	0.419672	0.412696
Chr14	795,050		
Chr21	333,660	0.441038	0.433978
Chr15	756,533		

Trisomy 18

Table 5 provides the calculated dose for chromosome 18 in a test sample (#11390). The calculated threshold for the positive diagnosis of T18 aneuploidy was set at 2 standard deviations from the mean of the qualified (normal) samples. A diagnosis for T18 was given based on the

chromosome dose in the test sample being greater than the set threshold. Chromosome 8 was used as the normalizing chromosome. In this instance chromosome 8 had the lowest variability and the greatest differentiability. Eight T18 samples were identified using chromosome doses, and were confirmed to be T18 by karyotype.

These data show that a normalizing chromosome can have both the lowest variability and the greatest differentiability.

TABLE 5

Chromosome Dose for a T18 aneuploidy (sample #11390, 47 XY +18)

Chromosome	Sequence Tag Density	Chromosome Dose for Chr 18	Threshold
Chr18	602,506	0.585069	0.530867
Chr8	1,029,803		

Trisomy 13

Table 6 provides the calculated dose for chromosome 13 in a test sample (#51236). The calculated threshold for the positive diagnosis of T13 aneuploidy was set at 2 standard deviations from the mean of the qualified samples. A diagnosis for T13 was given based on the chromosome dose in the test sample being greater than the set threshold. The chromosome dose for chromosome 13 was calculated using either chromosome 5 or the group of chromosomes 3, 4, 5, and 6 as the normalizing chromosome. One T13 sample was identified.

TABLE 6

Chromosome Dose for a T13 aneuploidy (sample #51236, 47 XY +13)

Chromosome	Sequence Tag Density	Chromosome Dose for Chr 13	Threshold
Chr13	692,242	0.541343	0.52594
Chr5	1,278,749		
Chr13	692,242	0.530472	0.513647
Chr3-6 [average]	1,304,954		

The sequence tag density for chromosomes 3-6 is the average tag counts for chromosomes 3-6.

The data show that the combination of chromosomes 3, 4, 5 and 6 provide a variability that is lower than that of chromosome 5, and the greatest differentiability than any of the other chromosomes.

Thus, a group of chromosomes can be used as the normalizing chromosome to determine chromosome doses and identify aneuploidies.

Turner Syndrome (monosomy X)

Table 7 provides the calculated dose for chromosomes X and Y in a test sample (#51238). The calculated threshold for the positive diagnosis of Turner Syndrome (monosomy X) was set for the X chromosome at < -2 standard deviations from the mean, and for the absence of the Y chromosome at < -2 standard deviations from the mean for qualified (normal) samples.

TABLE 7

Chromosome Dose for a Turners (XO) aneuploidy (sample #51238, 45 X)

Chromosome	Sequence Tag Density	Chromosome Dose for Chr X and Chr Y	Threshold
ChrX	873,631	0.786642	0.803832
Chr4	1,110,582		
ChrY	1,321	0.001542101	0.00211208
Chr_Total (1-22, X) (Average)	856,623.6		

A sample having an X chromosome dose less than that of the set threshold was identified as having less than one X chromosome. The same sample was determined to have a Y chromosome dose that was less than the set threshold, indicating that the sample did not have a Y chromosome. Thus, the combination of chromosome doses for X and Y were used to identify the Turner Syndrome (monosomy X) samples.

Thus, the method provided enables for the determination of CNV of chromosomes. In particular, the method enables for the determination of over- and under-representation chromosomal aneuploidies by massively parallel sequencing of maternal plasma cfDNA and identification of normalizing chromosomes for the statistical analysis of the sequencing data. The sensitivity and reliability of the method allow for accurate first and second trimester aneuploidy testing.

EXAMPLE 4

Determination of partial aneuploidy

The use of sequence doses was applied for assessing partial aneuploidy in a biological test sample of cfDNA that was prepared from blood plasma, and sequenced as described in Example 1. The sample was confirmed by karyotyping to have been derived from a subject with a partial deletion of chromosome 11.

Analysis of the sequencing data for the partial aneuploidy (partial deletion of chromosome 11 *i.e.* q21-q23) was performed as described for the chromosomal aneuploidies in the previous examples. Mapping of the sequence tags to chromosome 11 in a test sample revealed a noticeable loss of tag counts between base pairs 81000082-103000103 in the q arm of the chromosome relative to the tag counts obtained for corresponding sequence on chromosome 11 in the qualified samples (data not shown). Sequence tags mapped to the sequence of interest on chromosome 11 (810000082-103000103bp) in each of the qualified samples, and sequence tags mapped to all 20 megabase segments in the entire genome in the qualified samples *i.e.* qualified sequence tag densities, were used to determine qualified sequence doses as ratios of tag densities in all qualified samples. The average sequence dose, standard deviation, and coefficient of variation were calculated for all 20 megabase segments in the entire genome, and the 20-megabase sequence having the least variability was the identified normalizing sequence on chromosome 5 (13000014-33000033bp) (See Table 8), which was used to calculate the dose for the sequence of interest in the test sample (see Table 9). Table 8 provides the sequence dose for the sequence of interest on chromosome 11 (810000082-103000103bp) in the test sample that was calculated as the ratio of sequence tags mapped to the sequence of interest and the sequence tags mapped to the identified normalizing sequence. **Figure 10** shows the sequence doses for the sequence of interest in the 7 qualified samples (O) and the sequence dose for the corresponding sequence in the test sample (◊). The mean is shown by the solid line, and the calculated threshold for the positive diagnosis of partial aneuploidy that was set 5 standard deviations from the mean is shown by the dashed line. A diagnosis for partial aneuploidy was based on the sequence dose in the test sample being less than the set threshold. The test sample was verified by karyotyping to have deletion q21-q23 on chromosome 11.

Therefore, in addition to identifying chromosomal aneuploidies, the method of the invention can be used to identify partial aneuploidies.

TABLE 8

Qualified Normalizing Sequence, Dose and Variance for Sequence Chr11: 81000082-103000103 (qualified samples n=7)

	Chr11: 81000082-103000103		
	Avg	**Stdev**	**CV**
Chr5: 13000014-33000033	1.164702	0.004914	0.42

TABLE 9

Sequence Dose for Sequence of Interest (81000082-103000103) on Chromosome 11 (test sample 11206)

Chromosome Segment	Sequence Tag Density	Chromosome Segment Dose for Chr 11 (q21-q23)	Threshold
Chr11: 81000082-103000103	27,052	1.0434313	1.1401347
Chr5: 13000014-33000033	25,926		

Example 5

Demonstration of Detection of Aneuploidy

Sequencing data obtained for the samples described in Examples 2 and 3, and shown in figures 2-6 were further analyzed to illustrate the sensitivity of the method in successfully identifying aneuploidies in maternal samples. Normalized chromosome doses for chromosomes 21, 18, 13 X and Y were analyzed as a distribution relative to the standard deviation of the mean (Y-axis) and shown in Figure 11. The normalizing chromosome used is shown as the denominator (X-axis).

Figure 11 (A) shows the distribution of chromosome doses relative to the standard deviation from the mean for chromosome 21 dose in the unaffected samples (o) and the trisomy 21 samples (T21; Δ) when using chromosome 14 as the normalizing chromosome for chromosome 21. Figure 11 (B) shows the distribution of chromosome doses relative to the standard deviation from the mean for chromosome 18 dose in the unaffected samples (o) and the trisomy 18 samples (T18; Δ) when using chromosome 8 as the normalizing chromosome for chromosome 18. Figure 11 (C) shows the distribution of chromosome doses relative to the standard deviation from the mean for chromosome 13 dose in the unaffected samples (o) and the trisomy 13 samples (T13; Δ), using the average sequence tag density of the group of chromosomes 3, 4, 5, and 6 as the normalizing chromosome to

determine the chromosome dose for chromosome 13. Figure 11 (D) shows the distribution of chromosome doses relative to the standard deviation from the mean for chromosome X dose in the unaffected female samples (o), the unaffected male samples (Δ), and the monosomy X samples (XO; +) when using chromosome 4 as the normalizing chromosome for chromosome X. Figure 11 (E) shows the distribution of chromosome doses relative to the standard deviation from the mean for chromosome Y dose in the unaffected male samples (o the unaffected female sample s (Δ), and the monosomy X samples (+), when using the average sequence tag density of the group of chromosomes 1-22 and X as the normalizing chromosome to determine the chromosome dose for chromosome Y.

The data show that trisomy 21, trisomy 18, trisomy 13 were clearly distinguishable from the unaffected (normal) samples. The monosomy X samples were easily identifiable as having chromosome X dose that were clearly lower than those of unaffected female samples (Figure 11 (D)), and as having chromosome Y doses that were clearly lower than that of the unaffected male samples (Figure 11 (E)).

Therefore the method provided is sensitive and specific for determining the presence or absence of chromosomal aneuploidies in a maternal blood sample.

Example 6

Determination of Fetal Chromosomal Abnormalities Using Massively Parallel DNA Sequencing of Cell Free Fetal DNA from Maternal Blood: Test Set 1 Independent of Training Set 1

The study was conducted by qualified site clinical research personnel at 13 US clinic locations between April 2009 and July 2010 under a human subject protocol approved by institutional review boards (IRBs) at each institution. Informed written consent was obtained from each subject prior to study participation. The protocol was designed to provide blood samples and clinical data to support development of noninvasive prenatal genetic diagnostic methods. Pregnant women, age 18 years or older were eligible for inclusion. For patients undergoing clinically indicated CVS or amniocentesis blood was collected prior to performance of the procedure, and results of fetal karyotype was also collected. Peripheral blood samples (two tubes or ~20 mL total) were drawn from all subjects in acid citrate dextrose (ACD) tubes (Becton Dickinson). All samples were de-identified and assigned an anonymous patient ID number. Blood samples were shipped overnight to the laboratory in temperature controlled shipping containers provided for the study. Time elapsed between blood draw and sample receipt was recorded as part of the sample accessioning.

Site research coordinators entered clinical data relevant to the patient's current pregnancy and history into study case report forms (CRFs) using the anonymous patient ID number. Cytogenetic analysis of fetal karyotype from invasive prenatal procedure samples was performed per local laboratories and the results were also recorded in study CRFs. All data obtained on CRFs were entered into a clinical database the laboratory. Cell free plasma was obtained from individual blood tubes utilizing at two-step centrifugation process within 24-48 hours of sample of venipuncture.

Plasma from a single blood tube was sufficient for sequencing analysis. Cell-free DNA was extracted from cell-free plasma by using QIAamp DNA Blood Mini kit (Qiagen) according to the manufacturer's instructions. Since the cell free DNA fragments are known to be approximately 170 base pairs (bp) in length (Fan *et al.,* Clin Chem 56:1279-1286 [2010]) no fragmentation of the DNA was required prior to sequencing.

For the training set samples, cfDNA was sent to Prognosys Biosciences, Inc. (La Jolla, CA) for sequencing library preparation (cfDNA blunt ended and ligated to universal adapters) and sequencing using standard manufacturer protocols with the Illumina Genome Analyzer IIx instrumentation (http://www.illumina.com/). Single-end reads of 36 base pairs were obtained. Upon completion of the sequencing, all base call files were collected and analyzed. For the test set samples, sequencing libraries were prepared and sequencing carried out on Illumina Genome Analyzer IIx instrument. Sequencing library preparation was performed as follows. The full-length protocol described is essentially the standard protocol provided by Illumina, and only differs from the Illumina protocol in the purification of the amplified library: the Illumina protocol instructs that the amplified library be purified using gel electrophoresis, while the protocol described herein uses magnetic beads for the same purification step. Approximately 2 ng of purified cfDNA that had been extracted from maternal plasma was used to prepare a primary sequencing library using NEBNext™ DNA Sample Prep DNA Reagent Set 1 (Part No. E6000L; New England Biolabs, Ipswich, MA) for Illumina® essentially according to the manufacturer's instructions. All steps except for the final purification of the adaptor-ligated products, which was performed using Agencourt magnetic beads and reagents instead of the purification column, were performed according to the protocol accompanying the NEBNext™ Reagents for Sample Preparation for a genomic DNA library that is sequenced using the Illumina® GAII. The NEBNext™ protocol essentially follows that provided by Illumina, which is available at grcf.jhml.edu/hts/protocols/11257047_ChIP_Sample_Prep.pdf.

The overhangs of approximately 2 ng purified cfDNA fragments contained in 40µl were converted into phosphorylated blunt ends according to the NEBNext® End Repair Module by incubating the 40µl cfDNA with 5µl 10X phosphorylation buffer, 2µl deoxynucleotide solution mix (10 mM each dNTP), 1µl of a 1:5 dilution of DNA Polymerase I, 1µl T4 DNA Polymerase and 1µl T4 Polynucleotide Kinase provided in the NEBNext™ DNA Sample Prep DNA Reagent Set 1 in a 200 µl microfuge tube in a thermal cycler for 30 minutes at 20°C. The sample was cooled to 4°C, and purified using a QIAQuick column provided in the QIAQuick PCR Purification Kit (QIAGEN Inc., Valencia, CA) as follows. The 50µl reaction was transferred to 1.5 ml microfuge tube, and 250µl of Qiagen Buffer PB were added. The resulting 300µl were transferred to a QIAquick column, which was centrifuged at 13,000RPM for 1 minute in a microfuge. The column was washed with 750µl Qiagen Buffer PE, and re-centrifuged. Residual ethanol was removed by an additional centrifugation for 5 minutes at 13,000 RPM. The DNA was eluted in 39µl Qiagen Buffer EB by centrifugation. dA

tailing of 34µl of the blunt-ended DNA was accomplished using 16µl of the dA-tailing master mix containing the Klenow fragment (3' to 5' exo minus) (NEBNext™ DNA Sample Prep DNA Reagent Set 1), and incubating for 30 minutes at 37°C according to the manufacturer's NEBNext® dA-Tailing Module. The sample was cooled to 4°C, and purified using a column provided in the MinElute PCR Purification Kit (QIAGEN Inc., Valencia, CA) as follows. The 50µl reaction was transferred to 1.5 ml microfuge tube, and 250µl of Qiagen Buffer PB were added. The 300µl were transferred to the MinElute column, which was centrifuged at 13,000RPM for 1 minute in a microfuge. The column was washed with 750µl Qiagen Buffer PE, and re-centrifuged. Residual ethanol was removed by an additional centrifugation for 5 minutes at 13,000 RPM. The DNA was eluted in 15µl Qiagen Buffer EB by centrifugation. Ten microliters of the DNA eluate were incubated with 1µl of a 1:5 dilution of the Illumina Genomic Adapter Oligo Mix (Part No. 1000521), 15 µl of 2X Quick Ligation Reaction Buffer, and 4 µl Quick T4 DNA Ligase, for 15 minutes at 25°C according to the NEBNext® Quick Ligation Module. The sample was cooled to 4°C, and purified using a MinElute column as follows. One hundred and fifty microliters of Qiagen Buffer PE were added to the 30 µl reaction, and the entire volume was transferred to a MinElute column were transferred to a MinElute column, which was centrifuged at 13,000RPM for 1 minute in a microfuge. The column was washed with 750µl Qiagen Buffer PE, and re-centrifuged. Residual ethanol was removed by an additional centrifugation for 5 minutes at 13,000 RPM. The DNA was eluted in 28µl Qiagen Buffer EB by centrifugation. Twenty three microliters of the adaptor-ligated DNA eluate were subjected to 18 cycles of PCR (98°C for 30 seconds; 18 cycles of 98°C for 10 seconds, 65°C for 30 seconds, and 72°C for 30; final extension at 72°C for 5 minutes, and hold at 4°C) using Illumina Genomic PCR Primers (Part Nos. 100537 and 1000538) and the Phusion HF PCR Master Mix provided in the NEBNext™ DNA Sample Prep DNA Reagent Set 1, according to the manufacturer's instructions. The amplified product was purified using the Agencourt AMPure XP PCR purification system (Agencourt Bioscience Corporation, Beverly, MA) according to the manufacturer's instructions available at www.beckmangenomics.com/products/AMPureXPProtocol_000387v001.pdf. The Agencourt AMPure XP PCR purification system removes unincorporated dNTPs, primers, primer dimers, salts and other contaminates, and recovers amplicons greater than 100bp. The purified amplified product was eluted from the Agencourt beads in 40µl of Qiagen EB Buffer and the size distribution of the libraries was analyzed using the Agilent DNA 1000 Kit for the 2100 Bioanalyzer (Agilent technologies Inc., Santa Clara, CA). For both the training and test sample sets, single-end reads of 36 base pairs were sequenced.

Data Analysis and Sample Classification

Sequence reads 36 bases in length were aligned to the human genome assembly hg18 obtained from the UCSC database (http://hgdownload.cse.ucsc.edu/goldenPath/hg18/bigZips/). Alignments were carried out utilizing the Bowtie short read aligner (version 0.12.5) allowing for up to

two base mismatches during alignment (Langmead *et al.,* Genome Biol 10:R25 [2009]. Only reads that unambiguously mapped to a single genomic location were included. Genomic sites where reads mapped were counted and included in the calculation of chromosome doses (see below). Regions on the Y chromosome where sequence tags from male and female fetuses map without any discrimination were excluded from the analysis (specifically, from base 0 to base 2 x 10^6; base 10 x 10^6 to base 13 x 10^6; and base 23 x 10^6 to the end of chromosome Y).

Intra-run and inter-run sequencing variation in the chromosomal distribution of sequence reads can obscure the effects of fetal aneuploidy on the distribution of mapped sequence sites. To correct for such variation, a chromosome dose was calculated as the count of mapped sites for a given chromosome of interest is normalized to counts observed on a predetermined normalizing chromosome sequence. As described previously, a normalized chromosome sequence can be composed of a single chromosome or a group of chromosomes. The normalizing chromosome sequence was first identified in a subset of samples in the training set of samples that were unaffected *i.e.* qualified samples having diploid karyotypes for chromosomes of interest 21, 18, 13 and X, considering each autosome as a potential denominator in a ratio of counts with our chromosomes of interest. Denominator chromosomes *i.e.* normalizing chromosome sequences were selected that minimized the variation of the chromosome doses within and between sequencing runs. Each chromosome of interest was determined to have a distinct normalizing chromosome sequence (denominator) (Table 10). No single chromosome could be identified as a normalizing chromosome sequence for chromosome 13 as no one chromosome was determined to reduce the variability in the dose of chromosome 13 across samples *i.e.* the spread of the NCV values for chromosome 13 was not reduced sufficiently to allow for a correct identification of a T13 aneuploidy. Chromosomes 2-6 were chosen randomly and tested for their ability as a group to mimic the behavior of chromosome 13. The group of chromosomes 2-6 was found to diminish substantially the variability in the dose for chromosome 13 in the training samples, and was thus chosen as the normalizing chromosome sequence for chromosome 13. As described above, the variability in chromosome dose for chromosome Y is greater than 30 independently of which single chromosome is used as the normalizing chromosome sequence in determining the chromosome Y dose. The group of chromosomes 2-6 was found to diminish substantially the variability in the dose for chromosome Y in the training samples, and was thus chosen as the normalizing chromosome sequence for chromosome Y.

The chromosome doses for each of the chromosomes of interest in the qualified samples provides a measure of the variation in the total number of mapped sequence tags for each chromosome of interest relative to that of each of the remaining chromosomes. Thus, qualified chromosome doses can identify the chromosome or a group of chromosomes *i.e.* normalizing chromosome sequence that has a variation among samples that is closest to the variation of the

chromosome of interest, and that would serve as ideal sequences for normalizing values for further statistical evaluation.

Chromosome doses for all samples in the training set *i.e.* qualified and affected, also serve as the basis for determining threshold values when identifying aneuploidies in test samples as described in the following.

TABLE 10

Normalizing Chromosome Sequences for Determining Chromosome Doses

Chromosome of Interest	**Chromosome of Interest - Numerator (Chr mapped counts)**	**Normalizing Chromosome Sequence – Denominator (Chr mapped counts)**
21	Chr 21	Chr 9
18	Chr 18	Chr 8
13	Chr 13	Sum(Chr 2-6)
X	Chr X	Chr 6
Y	Chr Y	Sum(Chr 2-6)

For each chromosome of interest in each sample in the test set, a normalizing value was determined and used to determine the presence or absence of an aneuploidy. The normalizing value was calculated as a chromosome dose that can be further computed to provide a normalized chromosome value (NCV).

Chromosome doses

For the test set, a chromosome dose was calculated for each chromosome of interest, 21, 18, 13, X and Y for every sample. As provided in Table 10 above, the chromosome dose for chromosome 21 was calculated as a ratio of the number of tags in the test sample that mapped to chromosome 21 in the test sample, and the number of tags in the test sample that mapped to chromosome 9; the chromosome dose for chromosome 18 was calculated as a ratio of the number of tags in the test sample that mapped to chromosome 18 in the test sample, and the number of tags in the test sample that mapped to chromosome 8; the chromosome dose for chromosome 13 was calculated as a ratio of the number of tags in the test sample that mapped to chromosome 13 in the test sample, and the number of tags in the test sample that mapped to chromosomes 2-6; the chromosome dose for chromosome X was calculated as a ratio of the number of tags in the test sample that mapped to chromosome X in the test sample, and the number of tags in the test sample that mapped to chromosome 6; and the chromosome dose for chromosome Y was calculated as a ratio of the number of tags in the test sample that mapped to chromosome Y in the test sample, and the number of tags in the test sample that mapped to chromosomes 2-6.

Normalized chromosome values

Using the chromosome dose for each of the chromosomes of interest in each of the test samples, and the mean of the corresponding chromosome dose determined in the qualified samples of the training set, a normalized chromosome value (NCV) was calculated using the equation:

$$NCV_{ij} = \frac{x_{ij} - \hat{\mu}_j}{\hat{\sigma}_j}$$

where $\hat{\mu}_j$ AND $\hat{\sigma}_j$ are the estimated training set mean and standard deviation respectively for the *j-th* chromosome dose, and x_{ij} is the observed j-th chromosome dose for sample i. When chromosome doses are normally distributed, the NCV is equivalent to a statistical z-score for the doses. No significant departure from linearity is observed in a quantile-quantile plot of the NCVs from unaffected samples. In addition, standard tests of normality for the NCVs fail to reject the null hypothesis of normality.

For the test set, an NCV was calculated for each chromosome of interest, 21, 18, 13, X and Y for every sample. To insure a safe and effective classification scheme, conservative boundaries were chosen for aneuploidy classification. For classification of the autosomes' aneuploidy state, a NCV > 4.0 was required to classify the chromosome as affected (i.e. aneuploid for that chromosome) and a NCV < 2.5 to classify a chromosome as unaffected. Samples with autosomes that have an NCV between 2.5 and 4.0 were classified as "no call".

Sex chromosome classification in the test was performed by sequential application of NCVs for both X and Y as follows:

1. If NCV Y > -2.0 standard deviations from the mean of male samples, then the sample was classified as male (XY).
2. If NCV Y < -2.0 standard deviations from the mean of male samples, and NCV X > -2.0 standard deviations from the mean of female samples, then the sample was classified as female (XX).
3. If NCV Y < -2.0 standard deviations from the mean of male samples, and NCV X < -3.0 standard deviations from the mean of female samples, then the sample was classified as monosomy X, i.e. Turner syndrome.
4. If the NCVs did not fit into any of the above criteria, then the sample was classified as a "no call" for sex.

Results

Study Population Demographics

A total of 1,014 patients were enrolled between April 2009 and July 2010. The patient demographics, invasive procedure type and karyotype results are summarized in Table 11. The average age of study participants was 35.6 yrs (range 17 to 47 yrs) and gestational age ranged between 6 weeks, 1 day to 38 weeks, 1 day (mean 15 weeks, 4 days). The overall incidence of

abnormal fetal chromosome karyotypes was 6.8% with T21 incidence of 2.5%. Of 946 subjects with singleton pregnancies and karyotype, 906 (96%) showed at least one clinically recognized risk factor for fetal aneuploidy prior to prenatal procedure. Even eliminating those with advanced maternal age as their sole indication, the data demonstrates a very high false positive rate for current screening modalities. Ultrasound findings of increased nuchal translucency, cystic hygroma, or other structural congenital abnormality by ultrasound were most predictive of abnormal karyotype in this cohort.

TABLE 11

Patient Demographics

	Total Enrolled (N=1014)	**Training Set** (N=71)	**Test Set** (N=48)
Dates of Enrollment	Apr 2009 – Jul 2010	Apr 2009-Dec 2009	Jan 2010-Jun 2010
Number enrolled	1014	435	575
Maternal Age, yrs			
Mean (SD)	35.6 (5.66)	36.4 (6.05)	34.2 (8.22)
Min/Max	17/47	20/46	18/46
Not Specified, N	11	3	0
Ethnicity, N (%)			
Caucasian	636 (62.7)	50 (70.4)	24 (50.0)
Hispanic	167 (16.5)	6 (8.5)	13 (27.0)
Asian	63 (6.2)	6 (8.5)	5 (10.4)
Multi, more than one	53 (5.2)	6 (8.5)	1 (2.1)
African American	41 (4.0)	1 (1.3)	3 (6.3)
Other	36 (3.6)	2 (2.8)	1 (2.1)
Native American	9 (0.9)	0 (0.0)	1 (2.1)
Not Specified	9 (0.9)	0 (0.0)	0 (0.0)
Gestational Age, wks, days	15w 4d	14w 5d	15w 3d
Mean	6 w 1d/38w 1d	10 w 0d/23 w 1d	10w 4d/28 w 3d
Min/Max			
Number of Fetus, N			
1	982	67	47
2	30	4	1
3	2	0	0
Prenatal Procedure, N (%)	430 (42.4)	38 (53.5)	28 (58.3)

CVS	571 (56.3)	32 (45.1)	20 (41.7)
Amniocentesis	3 (0.3)	1 (1.4)	0 (0.0)
Not specified	10 (1.0)	0 (0.0)	0 (0.0)
Not performed			
Fetal Karyotype, N (%)			
46 XX	453* (43.9)	22* (29.7)	7* (14.6)
46 XY	474* (45.9)	26* (35.1)	14 (29.2)
47, +21, both sexes	25* (2.4)	10* (13.5)	13 (27.1)
47, +18, both sexes	14 (1.4)	5 (6.8)	8 (16.7)
47, +13, both sexes	4 (0.4)	2 (2.7)	1 (2.1)
45, X	8 (0.8)	3 (4.1)	3 (6.3)
Complex, other	18* (1.7)	6 (8.1)	2 (4.2)
Karyotype not available	36 (3.5)	0 (0.0)	0 (0.0)
Prenatal Screening Risks for Karyotyped Singletons, N (%)	**Non-sequenced** N=834	**Analyzed Training** N=65	**Analyzed Test** N=47
AMA only (≥35 years)	445 (53.4)	21 (44.7)	
Screen positive (trisomy)**	149 (17.9)	27 (41.5)	9 (19.1)
Increased NT	35 (4.2)	18 (27.7)	5 (10.6)
Cystic Hygroma	12 (1.4)	3 (4.6)	4 (8.5)
Cardiac Defect	14 (1.7)	5 (7.7)	4 (8.5)
Other Congenital Abnormality	78 (9.4)	0 (0.0)	3 (6.4)
Other Maternal Risk	64 (7.7)	4 (6.2)	1 (2.1)
None specified	37 (4.4)	5 (7.7)	0 (0.0)
		3 (4.6)	

*Includes results of fetuses from multiple gestations, **Assessed and reported by clinicians

Abbreviations: AMA = Advanced Maternal Age, NT = nuchal translucency

The distribution of diverse ethnic backgrounds represented in this study population is also shown in Table 11. Overall, 63% of the patients in this study were Caucasian, 17% Hispanic, 6% Asian, 5% multi-ethnic, and 4% African American. It was noted that the ethnic diversity varied significantly from site to site. For example, one site enrolled 60% Hispanic and 26% Caucasian subjects while three clinics all located in the same state, enrolled no Hispanic subjects. As expected, there were no discernible differences observed in our results for different ethnicities.

Training Data Set 1

The training set study selected 71 samples from the initial sequential accumulation of 435 samples that were collected between April 2009 and December 2009. All subjects with affected

fetus' (abnormal karyotypes) in this first series of subjects were included for sequencing and a random selection and number of non-affected subjects with adequate sample and data. Clinical characteristics of the training set patients were consistent with the overall study demographics as shown in Table 11. The gestational age range of the samples in the training set ranged from 10 weeks, 0 days to 23 weeks 1 day. Thirty-eight underwent CVS, 32 underwent amniocentesis and 1 patient did not have the invasive procedure type specified (an unaffected karyotype 46, XY). 70% of the patients were Caucasian, 8.5% Hispanic, 8.5% Asian, and 8.5% multi-ethnic. Six sequenced samples were removed from this set for the purposes of training: 4 samples from subjects with twin gestations (further discussed below), 1 sample with T18 that was contaminated during preparation, and 1 sample with a fetal karyotype 69, XXX, leaving 65 samples for the training set.

The number of unique sequence sites (i.e. tags identified with unique sites in the genome) varied from 2.2M in the early phases of the training set study to 13.7M in the latter phases due to improvements in sequencing technology over time. In order to monitor for any potential shifts in the chromosome doses over this 6-fold range in unique sites, different unaffected samples were run at the beginning and end of the study. For the first 15 unaffected samples run, the average number of unique sites was 3.8M and the average chromosome doses for chromosome 21 and chromosome 18 were 0.314 and 0.528, respectively. For the last 15 unaffected samples run, the average number of unique sites was 10.7M and the average chromosome doses for chromosome 21 and chromosome 18 were 0.316 and 0.529, respectively. There was no statistical difference between the chromosome doses for chromosome 21 and chromosome 18 over the time of the training set study.

The training set NCVs for chromosomes 21, 18 and 13 are shown on **Figure 12**. The results shown in **Figure 12** are consistent with an assumption of normality in that roughly 99% of the diploid NCVs would fall within $\pm$ 2.5 standard deviations of the mean. Of this set of 65 samples, 8 samples with clinical karyotypes indicating T21 had NCVs ranging from 6 to 20. Four samples having clinical karyotypes indicative of fetal T18 had NCVs ranging from 3.3 to 12, and the two samples having karyotypes indicative of fetal trisomy 13 (T13) had NCVs of 2.6 and 4. The spread of the NCVs in affected samples is due to their dependence on the percentage of fetal cfDNA in the individual samples.

Similar to the autosomes, the means and standard deviations for the sex chromosomes were established in the training set. The sex chromosome thresholds allowed 100% identification of male and female fetuses in the training set.

Test Data Set 1

Having established chromosome doses means and standard deviations from the training set, a test set of 48 samples was selected from samples collected between January 2010 and June 2010 from 575 total samples. One of the samples from a twin gestation was removed from the final analysis leaving 47 samples in the test set. Personnel preparing samples for sequencing and operating the

equipment were blinded to the clinical karyotype information. The gestational age range was similar to that seen in the training set (Table 11). 58% of the invasive procedures were CVS, higher than that of the overall procedural demographics, but also similar to the training set. 50% of subjects were Caucasian, 27% Hispanic, 10.4% Asian and 6.3% African American.

In the test set, the number of unique sequence tags varied from approximately 13M to 26M. For unaffected samples, the chromosome doses for chromosome 21 and chromosome 18 were 0.313 and 0.527, respectively. The test set NCVs for chromosome 21, chromosome 18 and chromosome 13 are shown in **Figure 13** and the classifications are given in Table 12.

TABLE 12

Test Set Classification Data Test Set Classification Data

T21 classification				
Karyotype	Unaffected for T21	T21	No Call	
Unaffected for T21	34			
47,XX or XY+21		13		
T18 classification				
Karyotype	Unaffected for T18	T18	No Call	
Unaffected for T18	39			
47,XX or XY+18		8		
T13 classification				
Karyotype	Unaffected for T13	T13	No Call	
Unaffected for T13	46			
47,XX or XY+13			1	
Sex Chromosome Classification				
Karyotype	XY	XX	MX*	No Call
46, XY	24			
46, XX		18		1
45, X			2	1
Cplx	1			

*MX is monosomy in the X chromosome with no evidence of Y chromosome

In the test set, 13/13 subjects having clinical karyotypes that indicated fetal T21 were correctly identified having NCVs ranging from 5 to 14. Eight/eight subjects having karyotypes that indicated fetal T18 were correctly identified having NCVs ranging from 8.5 to 22. The single sample

having a karyotype classified as T13 in this test set was classified as a no call with an NCV of approximately 3.

For the test data set, all male samples were correctly identified including a sample with complex karyotype, 46,XY + marker chromosome (unidentifiable by cytogenetics) (Table 3). Nineteen of twenty female samples were correctly identified, and one female sample was categorized as a no call. For three samples in the test set with karyotype of 45,X, two of the three were correctly identified as monosomy X and 1 was classified as a no call (Table 12).

Twins

Four of the samples initially selected for the training set and one of the samples in the test set were from twin gestations. The thresholds being employed here could be confounded by the differing amount of cfDNA expected in the setting of a twin gestation. In the training set, the karyotype from one of the twin samples was monochorionic 47,XY+21. A second twin sample was fraternal and amniocentesis was carried out on each of the fetuses individually. In this twin gestation, one of the fetuses had a karyotype of 47,XY+21 while the other had a normal karyotype, 46,XX. In both of these cases the cell free classification based on the methods discussed above classified the sample as T21. The other two twin gestations in the training set were classified correctly as non-affected for T21 (all twins showed diploid karyotype for chromosome 21). For the twin gestation sample in the test set, karyotype was only established for Twin B (46,XX) and the algorithm correctly classified as non-affected for T21.

Conclusion

The data show that massively parallel sequencing can be used to determine a plurality abnormal fetal karyotypes from the blood of pregnant women. These data demonstrate that 100% correct classification of samples with trisomy 21 and trisomy 18 can be identified using independent test set data. Even in the case of fetuses with abnormal sex chromosome karyotypes, none of the samples were incorrectly classified with the algorithm of the method. Importantly, the algorithm also performed well in determining the presence of T21 in two sets of twin pregnancies having at least one affected fetus, which has never been shown previously. Furthermore, this study examined a variety of sequential samples from multiple centers representing not only the range of abnormal karyotypes that one is likely to witness in a commercial clinical setting, but showing the significance of accurately classifying pregnancies non-affected by common trisomies to address the unacceptably high false positive rates that remain in prenatal screening today. The data provide valuable insight into the vast capabilities of employing this method in the future. Analysis of subsets of the unique genomic sites showed increases in the variance consistent Poisson counting statistics.

The data build on the findings of Fan and Quake who demonstrated that the sensitivity of noninvasive prenatal determination of fetal aneuploidy from maternal plasma using massively parallel sequencing is only limited by the counting statistics (Fan and Quake, PLos One 5, e10439 [2010]).

Because sequencing information was collected across the entire genome, this method is capable of determining any aneuploidy or other copy number variation including insertions and deletions. The karyotype from one of the samples had a small deletion in chromosome 11 between q21 and q23 that was observed as a ~10% decrease in the relative number of tags in a 25 Mb region starting at q21 when the sequencing data was analyzed in 500 kbase bins. In addition, in the training set, three of the samples had complex sex karyotypes due to mosaicism in the cytogenetic analysis. These karyotypes were: i) 47,XXX[9]/45,X[6], ii) 45,X [3]/ 46, XY[17], and iii) 47,XXX[13]/45,X[7]. Sample ii, which showed some XY-containing cells was correctly classified as XY. Samples i (from CVS procedure) and iii (from amniocentesis), which both showed a mixture of XXX and X cells by cytogenetic analysis (consistent with mosaic Turner syndrome), were classified as a no call and monosomy X, respectively.

In testing the algorithm, another interesting data point was observed having an NCV between -5 and -6 for chromosome 21 for one sample from the test set (**Figure 13**). Although this sample was diploid in chromosome 21 by cytogenetics, the karyotype showed mosaicism with partial triploidy for chromosome 9; 47, XX + 9 [9]/46, XX [6]. Since chromosome 9 is used in the denominator to determine the chromosome dose for chromosome 21 (Table 10), this lowers the overall NCV value. The ability of the use of normalizing chromosomes to determine fetal trisomy 9 in this sample is evidenced by the results provided in Example 7 below.

The conclusion of Fan, *et al* regarding the sensitivity of these methods is only correct if the algorithms being utilized are able to account for any random or systematic biases introduced by the sequencing method. If the sequencing data is not properly normalized the resulting analysis will be inferior to the counting statistics. Chiu, *et al* noted in their recent paper that their measurement of chromosomes 18 and 13 using the massively parallel sequencing method was imprecise, and concluded that more research was necessary to apply the method to the determination of T18 and T13 (Chiu *et al.,* BMJ 342:c7401 [2011]). The method utilized in the Chiu, *et al* paper simply uses the number of sequence tags on the chromosome of interest, in their case chromosome 21, normalized by the total number of tags in the sequencing run. The challenge for this approach is that the distribution of tags on each chromosome can vary from sequencing run to sequencing run, and thus increases the overall variation of the aneuploidy determination metric. In order to compare the results of the Chiu algorithm to the chromosome doses used in this example, the test data for chromosomes 21 and 18 was reanalyzed using the method recommended by Chiu, *et al.* as shown in **Figure 14**. Overall, a compression in the range of NCV for each of the chromosomes 21 and 18 was observed as well as a decrease in the determination rate with 10/13 T21 and 5/8 of the T18 samples correctly identified from our test set utilizing an NCV threshold of 4.0 for aneuploidy classification.

Ehrich, et al also focused only on T21 and used the same algorithm as Chiu, et al., (Ehrich *et al.,* Am J Obstet Gynecol 204:205 e1-e11 [2011]). In addition, after observing a shift in their test set

z-score metric from the external reference data *i.e.* training set, they retrained on the test set to establish the classification boundaries. Although in principle this approach is feasible, in practice it would be challenging to decide how many samples are required to train and how often one would need to retrain to ensure that the classification boundaries are correct. One method of mitigating this issue is to include controls in every sequencing run that measure the baseline and calibrate for quantitative behavior.

The data obtained using the present method show that massively parallel sequencing is capable of determining multiple fetal chromosomal abnormalities from the plasma of pregnant women when the algorithm for normalizing the chromosome counting data is optimized. The present method for quantification not only minimizes random and systematic variations between sequencing runs, but also allows for effective classification of aneuploidies across the entire genome, most notably T21 and T18. Larger sample collections are required to test the algorithm for T13 determination. To this end, a prospective, blinded, multi-site clinical study to further demonstrate the diagnostic accuracy of the present method is being performed.

Example 7

Determination of the Presence or Absence of at least 5 Different Chromosomal Aneuploidies in All Chromosomes of Individual Test Samples

To demonstrate the capability of the method to determine the presence or absence of any chromosomal aneuploidy in each of a set of maternal test samples (test set 1; Example 6), systematically determined normalizing chromosome sequences were identified in unaffected samples of the training set (training set 1; Example 6), and used to calculate chromosome doses for all chromosomes in each of the test samples. Determination of the presence or absence of any one or more different complete fetal chromosomal aneuploidies in each of the test and training set samples was accomplished from sequencing information obtained from a single sequencing run on each individual sample.

Using the chromosome densities *i.e.* the number of sequence tags identified for each chromosome in each of the samples of the training set described in Example 6, a systematically determined normalizing chromosome sequence consisting of a single chromosome or a group of chromosomes was determined by calculating a single chromosome dose for each of chromosomes 1-22, X and Y. The systematically determined normalizing chromosome sequence for each of chromosomes 1-22, X, and Y was determined by systematically calculating chromosome doses for each chromosome using every possible combination of chromosomes as the numerator. For example, for chromosome 21 as the chromosome of interest, chromosome doses were calculated as a ratio of (i) the number of sequence tags obtained for chromosome 21(chromosome of interest) and (ii) the number of sequence tags obtained for each of the remaining chromosomes, and the sum of the number of tags obtained for all possible combinations of the remaining chromosomes (excluding

chromosome 21) *i.e.* 1, 2, 3, 4, 5, *etc.* up to 20, 21, 22, X, and Y; 1 + 2, 1+3, 1+4, 1+5, *etc.* up to 1+20, 1+22,1+X, and 1+Y;, 1+2+3, 1+2+4, 1+2 +5 *etc.* up to 1+2+20, 1+2+22, 1+2+X, and 1+2+Y; 1+3+4, 1+3+5, 1+3+6 *etc.* up to 1+3+20, 1+3+22, 1+3+X, and 1+3+Y; 1+2+3+4, 1+2+3+5, 1+2+3+6 *etc.* up to 1+2+3+20, 1+2+3+22, 1+2+3+X, and 1+2+3+Y; and so on such that all possible combinations of all of chromosomes 1-20, 22, X and Y were used as a normalizing chromosome sequence (numerator) to determine all possible chromosome doses for each chromosome of interest in each of the qualified (aneuploid) samples in the training set. Chromosome doses were determined in the same manner for chromosome 21 in all training samples, and the systematically determined normalizing chromosome sequence for chromosome 21 was determined as the single or group of chromosomes resulting in a dose for chromosome 21 having the smallest variability across all training samples. The same analysis was repeated to determine the single chromosome or combination of chromosomes that would serve as the systematically determined normalizing chromosome sequence for each of the remaining chromosomes including chromosomes 13, 18, X and Y *i.e.* all possible combinations of chromosomes were used to determine the normalizing sequence (single chromosome or a group of chromosomes) for all other chromosomes of interest 1-12, 14-17, 19-20, 22, X and Y, in all training samples. Thus, all chromosomes were treated as chromosomes of interest, and a systematically determined normalizing sequence was determined for each of all chromosomes in each of the unaffected samples in the training set. Table 13 provides the single or the group of chromosomes that were identified as the systematically determined normalizing sequence for each of chromosomes of interest 1-22, X, and Y. As highlighted by Table 13, for some chromosomes of interest, the systematically determined normalizing chromosome sequence was determined to be a single chromosome (*e.g.* when chromosome 4 is the chromosome of interest), and for other chromosomes of interest, the systematically determined normalizing chromosome sequence was determined to be a group of chromosomes (*e.g.* when chromosome 21 is the chromosome of interest).

TABLE 13

Systematically Determined Normalizing Chromosome Sequences for All Chromosomes

Chromosome of Interest	Systematically Determined Normalizing Sequence	Chromosome of Interest	Systematically determined Normalizing Sequence
1	6+10+14+15+17+20	**13**	4+5
2	3+6+8+9+10	**14**	1+3+5+6+10+19
3	2+4+5+6+12	**15**	1+14+20
4	5	**16**	14+17+19+20+22
5	4+6+8+14	**17**	15+19+22

6	3+4+5+12+14	**18**	2+3+5+7
7	4+5+8+14+19+20	**19**	22
8	2+5+7	**20**	10+16+17+22
9	3+4+8+10+17+19+20+22	**21**	4+14+16+20+22
10	2+14+15+17+20	**22**	19
11	5+10+14+20+22	**X**	4+8
12	1+2+3+5+6+19	**Y**	4+6

The mean, standard deviation (SD) and coefficient of variance (CV) for the systematically determined normalizing chromosome sequence determined for each of all chromosomes are given in Table 14.

TABLE 14

Mean, Standard Deviation and Coefficient of Variance for all systematically determined normalizing chromosome sequences

Chromosome of interest	Mean	SD	CV
1	0.36637	0.00266	0.72%
2	0.31580	0.00068	0.22%
3	0.21983	0.00055	0.18%
4	0.98191	0.02509	2.56%
5	0.30109	0.00076	0.25%
6	0.21621	0.00059	0.27%
7	0.21214	0.00044	0.21%
8	0.25562	0.00068	0.27%
9	0.12726	0.00034	0.27%
10	0.24471	0.00098	0.40%
11	0.26907	0.00098	0.36%
12	0.12358	0.00029	0.23%
13[a]	0.26023	0.00122	0.47%
14	0.09286	0.00028	0.30%
15	0.21568	0.00147	0.68%
16	0.25181	0.00134	0.53%
17	0.46000	0.00248	0.54%
18[a]	0.10100	0.00038	0.38%
19	1.43709	0.02899	2.02%
20	0.19967	0.00123	0.62%

21[a]	0.07851	0.00053	0.67%
22	0.69613	0.01391	2.00%
X[b]	0.46865	0.00279	0.68%
Y[b]	0.00028	0.00004	14.97%

[a]Excluding trisomies

[b]Female fetus

The variance in chromosome doses across all training samples as reflected by the value of the CV, substantiates the use of systematically determined normalizing chromosome sequences to provide a large signal-to-noise ratio and dynamic range, allowing for the determination of the aneuploidies to be made with high sensitivity and high specificity, as shown in the following.

To demonstrate the sensitivity and specificity of the method, chromosome doses for all chromosomes of interest 1-22, X and Y were determined in each of the samples in the training set, and in each of all samples in the test set described in Example 5 using the corresponding systematically determined normalizing chromosome sequences provided in Table 13 above.

Using the systematically determined normalizing chromosome sequence for each of the chromosomes of interest, the presence or absence of any chromosomal aneuploidy was determined in each of the samples in the training set , and in each of the test samples *i.e.* it was determined whether each sample contained a complete fetal chromosomal aneuploidy of chromosome 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16,17, 18, 19, 20, 21, 22, X, and Y. Sequence information *i.e.* the number of sequence tags, was obtained for all chromosomes in each of the samples in the training set, and in each of the test samples, and a single chromosome dose for each of the chromosomes in each of the training and test samples was calculated as described above using the number of sequence tags obtained for the systematically determined normalizing chromosome sequences corresponding to those determined in the trained set (Table 13). The number of sequence tags obtained in each of the training samples for the systematically determined normalizing chromosome sequences was used to determine the chromosome doses for each chromosome in each of the training samples, and the number of sequence tags obtained in each of the test samples for the systematically determined normalizing chromosome sequence was used to determine the chromosome dose for each chromosome for each of the test samples. To ensure safe and effective classification of aneuploidies, the same conservative boundaries were chosen as described in Example 6.

Training Set Results

A plot of the chromosome doses for chromosomes 21, 18 and 13 in the training set of samples using the systematically determined normalizing chromosome sequence is given in **Figure 15**. When using the systematically determined normalizing chromosome sequence *i.e.* the group of chromosomes 4+14+16+20+22, 8 samples with clinical karyotypes indicating T21 had NCVs between

5.4 and 21.5. When using the systematically determined normalizing chromosome sequence *i.e.* the group of chromosomes 2+3+5+7, 4 samples with clinical karyotypes indicating T18 had NCVs between 3.3 and 15.3. When using the systematically determined normalizing chromosome sequence *i.e.* the group of chromosomes 4+5, 2 samples with clinical karyotypes indicating T13 had NCVs of 8.0 and 12.4. The T21 samples of the training set are shown as the last 8 samples of the chromosome 21 data (O); theT18 samples of the training set are shown as the last 4 samples of the chromosome 18 data (Δ); and the T13 samples of the training set are shown as the last 2 samples of the chromosome 13 data (□).

These data show that normalizing chromosome sequences can be used to determine and correctly classify different complete fetal chromosomal aneuploidies with great confidence. Since all samples with affected karyotypes had NCVs greater than 3, there is less than approximately 0.1% probability that these samples are part of the unaffected distribution.

Similarly to the autosomes, when the systematically determined normalizing chromosome sequence (*i.e.* the group of chromosomes 4+8) was used for chromosome X, and when the systematically determined normalizing chromosome sequence (*i.e.* the group of chromosomes 4+6) was used for chromosome Y, all of the male and female fetuses in the training set were correctly identified. In addition, all 5 of the monosomy X samples were identified. **Figure 18A** shows a plot of NCVs determined for the X chromosome (X-axis) and NCVs determined for the Y chromosome (Y axis) for each of the samples in the training set. All of the samples which are monosomy e X by karyotype have NCV values of less than -4.83. Those monosomy X samples that have karyotypes consistent with a 45,X karyotype (full or mosaic) have a Y NCV value close to zero as expected. Female samples cluster around NCV = 0 for both X and Y.

Test Set Results

A plot of the chromosome doses for chromosomes 21, 18 and 13 in the test samples using the relevant systematically determined normalizing chromosome sequences is given in **Figure 16**. When using the systematically determined normalizing chromosome sequence (*i.e.* the group of chromosomes 4+14+16+20+22), then 13 of 13 samples with clinical karyotypes indicating T21 were correctly identified with NCVs between 7.2 and 16.3. When using the systematically determined normalizing chromosome sequence (*i.e.* the group of chromosomes 2+3+5+7), then all 8 samples with clinical karyotypes indicating T18 were identified with NCVs between 12.7 and 30.7. When using the systematically determined normalizing chromosome sequence (*i.e.* the group of chromosomes 4+5), then the only one sample with clinical karyotypes indicating T13 was correctly identified with an NCV of 8.6. The T21 samples of the test set are shown as the last 13 samples of the chromosome 21 data (O); theT18 samples of the test set are shown as the last 8 samples of the chromosome 18 data (Δ); and the T13 sample of the test set is shown as the last sample of the chromosome 13 data (□).

These data show that systematically determined normalizing chromosome sequences can be used to determine and correctly classify different complete fetal chromosomal aneuploidies with great confidence. Similar to the training set, all samples with affected karyotypes had NCVs greater than 7, which indicated an infinitesimally small probability that these samples are part of the unaffected distribution. (**Figure 16**).

Similarly to the autosomes, when the systematically determined normalizing chromosome sequence (*i.e.* the group of chromosomes 4+8) was used for chromosome X, and when the systematically determined normalizing chromosome sequence (*i.e.* the group of chromosomes 4+6) was used for chromosome Y, all of the male and female fetuses in the test set were correctly identified. In addition, all 3 of the monosomy X samples were determined. **Figure 18B** shows a plot of NCVs determined for the X chromosome (X-axis) and NCVs determined for the Y chromosome (Y axis) for each of the samples in the test set

As previously described, the present method allows for determining the presence or absence of a complete, or partial, chromosomal aneuploidy of each of chromosomes 1-22, X, and Y in each sample. In addition to determining complete chromosomal aneuploidies T13, T18, T21, and monosomy X, the method determined the presence of a trisomy of chromosome 9 in one of the test samples. When using the systematically determined normalizing chromosome sequence (*i.e.* the group of chromosomes 3+4+8+10+17+19+20+22), for chromosome of interest 9, a sample having an NCV of 14.4 was identified (**Figure 17**). This sample corresponded to the test sample in Example 6 that was suspected of being aneuploid for chromosome 9 following the calculation of an aberrantly low dose for chromosome 21 (for which chromosome 9 was used as the normalizing chromosome sequence in Example 6).

The data show that 100% of the samples having clinical karyotypes indicating T21, T13 T18, T9 and monosomy X were correctly identified. **Figure 19** shows a plot of the NCVs for each of chromosomes 1-22 in each of the 47 test samples. Medians of NCVs were normalized to zero. The data show that the method of the invention (including the use of systematically determined normalizing chromosome sequences) determined the presence of all 5 types of chromosomal aneuploidies that were present in this test set with 100% sensitivity and 100% specificity, and clearly indicate that the method can identify any complete chromosomal aneuploidy for any one of chromosomes 1-22, X, and Y, in any sample.

EXAMPLE 8

Determination of the presence or absence of a partial fetal chromosomal aneuploidy: determination of Cat eye syndrome

DiGeorge syndrome (22q11.2 deletion syndrome), a disorder caused by a defect in chromosome 22, results in the poor development of several body systems. Medical problems

commonly associated with DiGeorge syndrome include heart defects, poor immune system function, a cleft palate, poor function of the parathyroid glands and behavioral disorders. The number and severity of problems associated with DiGeorge syndrome vary greatly. Almost everyone with DiGeorge syndrome needs treatment from specialists in a variety of fields.

To determine the presence or absence of a partial deletion of fetal chromosome 22, a blood sample is obtained by venipuncture for the mother, and cfDNA is prepared as described in the Examples above. The purified cfDNA is ligated to adaptors and subjected to cluster amplification using the Illumina cBot cluster station. Massively parallel sequencing is performed using reversible dye terminators to generate millions of 36bp reads. The sequence reads are aligned to the human hg19 reference genome, and the reads that are uniquely mapped to the reference genome are counted as tags.

A set of qualified samples all known to be diploid for chromosome 22 *i.e.* chromosome 22 or any portion thereof is known to be present only in a diploid state, are first sequenced and analyzed to obtain a number of sequence tags for each of 1000 segments of 3 megabases (Mb) (excluding the region 22q11.2). Given that the human genome comprises approximately 3 billion bases (3Gb), the 1000 segments of 3 Mb each approximately composes the remainder of the genome. Each of the 1000 segments can serve individually or as in a group of segment sequences that are used to determine the normalizing segment sequence for the segment of interest *i.e.* the 3Mb region of 22q11.2. The number of sequence tags mapped to every single 1000bp segment is used individually to compute segment doses for the 3Mb region of 22q11.2. In addition, all possible combinations of two or more segments are used to determine segment doses for the segment of interest in all qualified samples. The single 3Mb segment or the combination of two or more 3Mb segments that result in the segment dose having the lowest variability across samples is chosen as the normalizing segment sequence.

The number of sequence tags mapped to the segment of interest in each of the qualified samples is used to determine a segment dose in each of the qualified samples. The mean and standard deviation of the segment doses in all qualified samples is calculated, and used to set threshold s to which segment doses determined in test samples can be compared. Preferably, normalized segment values (NSV) are calculated for all segments of interest in all qualified samples, and used to set the threshold values.

Subsequently, the number of tags mapped to the normalizing segment sequence in the corresponding test sample is used to determine the dose of the segment of interest in the test sample. A normalized segment value (NSV) is calculated for the segment in the test sample as described previously and the NCV of the segment of interest in the test sample is compared to the threshold determined using the qualified samples to determine the presence or absence of a deletion of 22q11.2 in the test sample.

A test NCV < -3, indicates that a loss in the segment of interest *i.e.* partial deletion of chromosome 22 (22q11.2) is present in the test sample.

EXAMPLE 9

Stool DNA Testing for Prediction of Outcome for StageII Colorectal Cancer Patients

Around 30% of all stage II colon cancer patients will relapse and die of their disease. Stage II colon cancers of patients who had relapse of disease showed significantly more losses on chromosomes 4, 5, 15q, 17q and 18q. In particular, stage II colon cancer patients losses on 4q22.1–4q35.2 have been shown to be associated with worse outcome. Determination of the presence or absence of these genomic alterations may aid in selecting patients for adjuvant therapy (Brosens *et al.,* Analytical Cellular Pathology / Cellular Oncology 33: 95–104 [2010]).

To determine the presence or absence of one or more chromosomal deletions in the 4q22.1 to 4q35.2 region in patients with stage II colorectal cancer, stool and/or plasma samples are obtained from the patient(s). Stool DNA is prepared according to the method described by Chen *et al.,* J Natl Cancer Inst 97:1124-1132 [2005]); and plasma DNA is prepared according to the method described in the Examples above. DNA is sequenced according to an NGS method described herein, and the sequence information for the patient(s) sample(s) is used to calculate segment doses for one or more segments spanning the 4q22.1 to 4q35.2 region. Segment doses are determined using normalizing segment sequences that are determined a priori by in a set of qualified stool and/or plasma samples, respectively. Segment doses in the test samples (patient samples) are calculated, and the presence or absence of one or more partial chromosomal deletions within the 4q22.1 to 4q35.2 region is determined by comparing the NSV for each of the segments of interest to the threshold set from the NSV in the set of qualified samples.

While preferred embodiments of the present invention have been shown and described herein, it will be obvious to those skilled in the art that such embodiments are provided by way of example only. Numerous variations, changes, and substitutions will now occur to those skilled in the art without departing from the invention. It should be understood that various alternatives to the embodiments of the invention described herein may be employed in practicing the invention. It is intended that the following claims define the scope of the invention and that methods and structures within the scope of these claims and their equivalents be covered thereby.

Claims

1. A method for determining the presence or absence of any twenty or more different complete fetal chromosomal aneuploidies in a maternal plasma test sample comprising a mixture of fetal and maternal cell-free DNA molecules, said method comprising:

(a) sequencing at least a portion of said cell-free DNA molecules to obtain sequence information for said fetal and maternal cell-free DNA molecules in said sample;

(b) using said sequence information to identify a number of sequence tags for each of any twenty or more chromosomes of interest selected from chromosomes 1-22, X, and Y and to identify a number of sequence tags for a normalizing chromosome for each of said twenty or more chromosomes of interest;

(c) using said number of sequence tags identified for each of said twenty or more chromosomes of interest and said number of sequence tags identified for each said normalizing chromosome to calculate a single chromosome dose for each of said twenty or more chromosomes of interest; and

(d) comparing each of said single chromosome doses for each of said twenty or more chromosomes of interest to a threshold value for each of said twenty or more chromosomes of interest, and thereby determining the presence or absence of any twenty or more different complete fetal chromosomal aneuploidies in said sample;

wherein each said normalizing chromosome for each of said twenty or more chromosomes of interest is determined by systematically calculating all possible chromosome doses for each of the twenty or more chromosomes of interest in a training set of samples comprising qualified samples comprising chromosomes that are present in known copy number and which are not aneuploid for each of the twenty or more chromosomes of interest, by using the number of sequence tags identified for each of the remaining chromosomes and the sum of the number of sequence tags identified for all possible combinations of the remaining chromosomes to determine which single chromosome or group of chromosomes results in the least variability of the chromosome dose for each of the twenty or more chromosomes of interest across said qualified samples.

2. The method of Claim 1, wherein step (c) comprises calculating a single chromosome dose for each of said chromosomes of interest as the ratio of the number of sequence tags identified for each of said chromosomes of interest and the number of sequence tags identified for said normalizing chromosome sequence for each of said chromosomes of interest.

3. The method of any one of Claims 1-2, wherein step (c) comprises:

(i) calculating a sequence tag density ratio for each of said chromosomes of interest, by relating the number of sequence tags identified for each of said chromosomes of interest in step (b) to the length of each of said chromosomes of interest;

(ii) calculating a sequence tag density ratio for each said normalizing chromosome by relating the number of sequence tags identified for said normalizing chromosome sequence in step (b) to the length of each said normalizing chromosome; and

(iii) using the sequence tag density ratios calculated in steps (i) and (ii) to calculate a single chromosome dose for each of said chromosomes of interest, wherein said chromosome dose is calculated as the ratio of the sequence tag density ratio for each of said chromosomes of interest and the sequence tag density ratio for said normalizing chromosome sequence for each of said chromosomes of interest.

4. The method of any one of Claims 1-3, wherein said any twenty or more chromosomes of interest selected from chromosomes 1-22, X, and Y is all of chromosomes 1-22, X, and Y, and wherein the presence or absence of complete fetal chromosomal aneuploidies of all of chromosomes 1-22, X, and Y is determined.

5. The method of any one of Claims 1-4, wherein said normalizing chromosome sequence is a single chromosome selected from chromosomes 1-22, X, and Y.

6. The method of any one of Claims 1-4, wherein said normalizing chromosome sequence is a group of chromosomes selected from chromosomes 1-22, X, and Y.

7. The method of any one of Claims 1-6, wherein said different complete chromosomal

aneuploidies are selected from complete chromosomal trisomies, complete chromosomal monosomies and complete chromosomal polysomies.

8. The method of any one of Claims 1-7, wherein said different complete fetal chromosomal aneuploidies are selected from trisomy 2, trisomy 8, trisomy 9, trisomy 21, trisomy 13, trisomy 16, trisomy 18, trisomy 22, 47,XXY, 47,XXX, 47,XYY, and monosomy X.

9. The method of any one of Claims 1-8, wherein steps (a)-(d) are repeated for test samples from different maternal subjects, and wherein the method comprises determining the presence or absence of any twenty or more different complete fetal chromosomal aneuploidies in each of said samples.

10. The method of any one of Claims 1-9, further comprising calculating a normalized chromosome value (NCV), wherein said NCV relates said chromosome dose to the mean of the corresponding chromosome dose in a set of qualified samples as:

$$NCV_{ij} = \frac{x_{ij} - \hat{\mu}_j}{\hat{\sigma}_j}$$

where $\hat{\mu}_j$ and $\hat{\sigma}_j$ are the estimated mean and standard deviation, respectively, for the j-th chromosome dose in a set of qualified samples, and x_{ij} is the observed j-th chromosome dose for test sample i.

11. The method of any one of the preceding Claims, wherein said test sample is a maternal sample selected from blood, plasma, serum, urine and saliva samples.

12. The method of any one of the preceding Claims, wherein said sequencing is next generation sequencing (NGS).

13. The method of any one of the preceding Claims, wherein said sequencing is massively parallel sequencing using sequencing-by-synthesis with reversible dye terminators.

14. The method of any one of the preceding Claims, wherein said sequencing is sequencing-by-ligation.

15. The method of any one of the preceding Claims, wherein said sequencing comprises an amplification.

16. The method of any one of the preceding Claims, wherein said sequencing is single molecule sequencing.

17. The method of any one of the preceding Claims, comprising, prior to step a), the step of preparing a sequencing library from said mixture of fetal and maternal nucleic acid molecules; wherein preparing said library comprises the consecutive steps of end-repairing, dA-tailing and adaptor ligating said nucleic acids, wherein said consecutive steps exclude purifying the end-repaired products prior to the dA-tailing step and exclude purifying the dA-tailing products prior to the adaptor-ligating step.

18. A method for determining the presence or absence of any four or more different complete fetal chromosomal aneuploidies for each of four or more chromosomes of interest in a maternal test sample comprising fetal and maternal nucleic acids, said method comprising:

(a) preparing a sequencing library from said mixture of fetal and maternal nucleic acid molecules; wherein preparing said library comprises the consecutive steps of end-repairing, dA-tailing and adaptor ligating said nucleic acids, wherein said consecutive steps exclude purifying the end-repaired products prior to the dA-tailing step and exclude purifying the dA-tailing products prior to the adaptor-ligating step;

(b) obtaining sequence information for said fetal and maternal nucleic acids in said sample;

(c) using said sequence information to identify a number of sequence tags for each of said any four or more chromosomes of interest, wherein the four or more chromosomes of interest are selected from chromosomes 1-22, X, and Y and to identify a number of sequence tags for a normalizing chromosome for each of said any four or more chromosomes of interest;

(d) using said number of sequence tags identified for each of said any four or more chromosomes of interest and said number of sequence tags identified for each said normalizing chromosome to calculate a single chromosome dose for each of said any four or more chromosomes of interest; and

(e) comparing each of said single chromosome doses for each of said any four or more chromosomes of interest to a threshold value for each of said four or more chromosomes of interest, and thereby determining the presence or absence of any four or more complete different fetal chromosomal aneuploidies in said sample;

wherein each said normalizing chromosome for each of said four or more chromosomes of interest is determined by systematically calculating all possible chromosome doses for each of the four or more chromosomes of interest in a training set of samples comprising qualified samples comprising chromosomes that are present in known copy number and which are not aneuploid for each of the four or more chromosomes of interest, by using the number of sequence tags identified for each of the remaining chromosomes and the sum of the number of sequence tags identified for all possible combinations of the remaining chromosomes to determine which single chromosome or group of chromosomes results in the least variability of the chromosome dose for each of the four or more chromosomes of interest across said qualified samples.

19. The method of claim 18, wherein said consecutive steps comprise heat inactivating the end-repair reaction mixture prior to the dA-tailing step, and heat inactivating the dA-tailing reaction mixture prior to the adaptor-ligating step.

20. The method of Claim 18, wherein step (d) comprises calculating a single chromosome dose for each of said chromosomes of interest as the ratio of the number of sequence tags identified for each of said chromosomes of interest and the number of sequence tags identified for said normalizing chromosome sequence for each of said chromosomes of interest.

21. The method of any one of Claims 18-20, wherein step (d) comprises:

(i) calculating a sequence tag density ratio for each of said chromosomes of interest, by relating the number of sequence tags identified for each of said chromosomes of interest in step (c) to the length of each of said chromosomes of interest;

(ii) calculating a sequence tag density ratio for each said normalizing chromosome by relating the number of sequence tags identified for said normalizing chromosome sequence in step (c) to the length of each said normalizing chromosome; and

(iii) using the sequence tag density ratios calculated in steps (i) and (ii) to calculate a single chromosome dose for each of said chromosomes of interest, wherein said chromosome dose is calculated as the ratio of the sequence tag density ratio for each of said chromosomes of interest and the sequence tag density ratio for said normalizing chromosome sequence for each of said chromosomes of interest.

22. The method of any one of Claims 18-21, wherein said any four or more chromosomes of interest selected from chromosomes 1-22, X, and Y comprise at least twenty chromosomes selected from chromosomes 1-22, X, and Y, and wherein the presence or absence of at least twenty different complete fetal chromosomal aneuploidies is determined.

23. The method of any one of Claims 18-22, wherein said any four or more chromosomes of interest selected from chromosomes 1-22, X, and Y is all of chromosomes 1-22, X, and Y, and wherein the presence or absence of complete fetal chromosomal aneuploidies of all of chromosomes 1-22, X, and Y is determined.

24. The method of any one of Claims 18-23, wherein said normalizing chromosome sequence is a single chromosome selected from chromosomes 1-22, X, and Y.

25. The method of any one of Claims 18-24, wherein said normalizing chromosome sequence is a group of chromosomes selected from chromosomes 1-22, X, and Y.

26. A method for determining a fetal chromosomal aneuploidy in a maternal sample comprising a mixture of fetal and maternal nucleic acids molecules, said method comprising:

(a) preparing a sequencing library from said mixture of fetal and maternal nucleic acid molecules; wherein preparing said library comprises the consecutive steps of end-repairing, dA-tailing and adaptor ligating said nucleic acids, wherein said consecutive steps exclude purifying the end-repaired products prior to the dA-tailing step and exclude purifying the dA-tailing products prior to the adaptor-ligating step;

(b) sequencing at least a portion of said nucleic acid molecules in said sequencing library, thereby obtaining sequence information for a plurality of fetal and maternal nucleic acid molecules of a maternal sample; and

(c) using said sequence information to identify a number of sequence tags for each of any one or more chromosomes of interest, wherein the one or more chromosomes of interest are selected from chromosomes 1-22, X and Y and to identify a number of sequence tags for a normalizing segment sequence for each of said any one or more chromosomes of interest;

(d) using said number of sequence tags identified for each of said any one or more chromosomes of interest and said number of sequence tags identified for said normalizing segment sequence to calculate a single chromosome dose for each of said any one or more chromosomes of interest; and

(e) comparing each of said single chromosome doses for each of said any one or more chromosomes of interest to a threshold value for each of said one or more chromosomes of interest, and thereby determining the presence or absence of one or more different complete fetal chromosomal aneuploidies in said sample;

wherein each said normalizing chromosome segment for each of said one or more chromosomes of interest is determined by systematically calculating all possible chromosome doses for each of the one or more chromosomes of interest in a training set of samples comprising qualified samples comprising chromosomes that are present in known copy number and which are not aneuploid for each of the one or more chromosomes of interest, by using the number of sequence tags identified for segments of each of the remaining chromosomes and the sum of the number of sequence tags identified for all possible combinations of segments of the remaining chromosomes to determine which single segment or group of segments results in the least variability of the chromosome dose for each of the one or more chromosomes of interest across said qualified samples.

27. The method of claim 26, wherein said consecutive steps comprise heat inactivating the end-repair reaction mixture prior to the dA-tailing step, and heat inactivating the dA-tailing reaction mixture prior to the adaptor-ligating step.

28. A method for determining the presence or absence of different partial fetal chromosomal aneuploidies for one or more segments of any one or more chromosomes of interest in a maternal test sample comprising fetal and maternal nucleic acids, said method comprising:

(a) preparing a sequencing library from said mixture of fetal and maternal nucleic acid molecules; wherein preparing said library comprises the consecutive steps of end-repairing, dA-tailing and adaptor ligating said nucleic acids, wherein said consecutive steps exclude purifying the end-repaired products prior to the dA-tailing step and exclude purifying the dA-tailing products prior to the adaptor-ligating step;

(b) sequencing at least a portion of said nucleic acid molecules in said sequencing library, thereby obtaining sequence information for said fetal and maternal nucleic acid molecules in said sample;

(c) using said sequence information to identify a number of sequence tags for each of any one or more segments of any one or more chromosomes of interest, wherein the one or more segments are selected from chromosomes 1-22, X, and Y and to identify a number of sequence tags for a normalizing segment sequence for each of said any one or more segments of any one or more chromosomes of interest;

(d) using said number of sequence tags identified for each of said any one or more segments of any one or more chromosomes of interest and said number of sequence tags identified for said normalizing segment sequence to calculate a single chromosome dose for each of said any one or more segments of any one or more chromosomes of interest; and

(e) comparing each of said single segment doses for each of said any one or more segments of any one or more chromosomes of interest to a threshold value for each of said any one or more chromosomal segments of any one or more chromosome of interest, and thereby determining the presence or absence of one or more different partial fetal chromosomal aneuploidies in said sample,

wherein each said normalizing chromosome segment for each of said one or more segments of any one or more chromosomes of interest is determined by systematically

calculating all possible chromosome doses for each of the one or more segments of any one or more chromosomes of interest in a training set of samples comprising qualified samples comprising chromosomes that are present in known copy number and which are not aneuploid for each of the one or more chromosomes of interest, by using the number of sequence tags identified for segments of each of the remaining chromosomes and the sum of the number of sequence tags identified for all possible combinations of the segments of the remaining chromosomes to determine which single segment or group of segments results in the least variability of the chromosome dose for each of the one or more segments of any one or more chromosomes of interest across said qualified samples.